Universidad de Málaga
Escuela Técnica Superior de Ingeniería de Telecomunicación

TESIS DOCTORAL

Bayesian modelling of fault diagnosis in mobile communication networks

Autor:

Raquel Barco Moreno

Directores:

Luis Díez del Río
Volker Wille

UNIVERSIDAD DE MÁLAGA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

Reunido el tribunal examinador en el día de la fecha, constituido por:

Presidente: Dr. D. ____

Secretario: Dr. D. ____

Vocales: Dr. D. ____

Dr. D. ____

Dr. D. ____

para juzgar la Tesis Doctoral titulada *Bayesian modelling of fault diagnosis in mobile communication networks* realizada por Dª. Raquel Barco Moreno y dirigida por Dr. D. Luis Díez del Río y Dr. D. Volker Wille, acordó por

____ otorgar la calificación de

____ y para que conste, se

extiende firmada por los componentes del tribunal la presente diligencia.

Málaga a ______ de __________ del ______

El presidente:

Fdo.: ____________________

El secretario:

Fdo.: ____________________

El vocal:

Fdo.: ____________

El vocal:

Fdo.: ____________

El vocal:

Fdo.: ____________

A mis padres

"Toutes les sciences sont tellement liées ensemble, qu'il est plus facile de les apprendre toutes à la fois, que d'en isoler une des autres". René Descartes (1596-1650)

Acknowledgements

This thesis has been an important part of my daily activities during more than six years. The thought "I should be working on my thesis" has constantly accompanied me during both working and spare time. Too frequently, the thesis has absorbed me, stealing time from my other activities, my family and friends. It is difficult to believe that now it is time to close this endless period of my life.

If I look back and evaluate the results of this thesis, I cannot avoid thinking that for me the most important achievement has been the colleagues, now friends, that I have met. All the lonely hours spent in this thesis along so many years are worthwhile because of those people. This is why I would like to thank those that have directly or indirectly supported me in these years.

The roots of this thesis lie in my experience at the European Space Agency in 1997. Amongst other tasks, I was assigned the design of an automated diagnosis system for satellite ground stations. Thanks to Ramón Segura, who was the one that first saw the necessity of such a system and who was my supervisor at the beginning of the work. Thanks to Klaus-Jürgen Schulz, the section head, who supported me to do the diagnosis study, despite taking up a lot of time from my other "operational" activities.

In 1999, I came back to Spain after two years in Germany. The change was quite frustrating at the beginning. I missed the team work and the meetings of my previous work compared to the more independent way of researching at university. At this point, thanks to Carlos Camacho, who (once more) gave me the opportunity to work at Nokia when a research center was open at Málaga. This was just what I needed at that time.

After some months, Nokia decided to start a project on automatic troubleshooting in cellular networks. Because of my previous experience, this project fitted perfectly with my interests and, at the same time, gave me the opportunity to start my thesis on that topic. On the one hand, there was a risk of Nokia closing the project (as actually happened some years later) letting my thesis unfinished. On the other hand, the time dedicated to my thesis was less than if I would have done a more conventional thesis because only part of the time I spent in Nokia could be used for my thesis. However, the experience of working in a team with staff in Spain, Denmark and UK was excellent. Thanks to all my colleagues in that project for the very good moments that we shared. Thanks to Sagar Patel, our link with operators, for being such a good friend and phoning me so often, in spite of living in UK. Thanks to Lars Moltsen, who was the first to propose Bayesian Networks, for doing work such an enjoyable activity. Thanks to Rafael Guerrero, my KAT man, for bearing me when I wrote such awful equations and specifications for the tool (and he even implemented them!). Thanks to Gustavo Hylander, Martti Partanen and Volker Wille, who were able to create a wonderful team spirit in which everyone had something to contribute with. I will never forget our meetings in Málaga. I owe this team much of this thesis.

Unfortunately, after three years, the Nokia center in Málaga was disbanded and with it the troubleshooting project disappeared. Some months later, I was invited to participate in a MICYT project, which was a support for this thesis. Thanks to all my colleagues in this project. Thanks to Matías Toril for his wise advice and fruitful cooperation. Thanks to Mariano Fernández, for doing the project so easy-going, for being so honest and for giving me the freedom to carry out any initiative I had on the project.

From 2004, I had the chance to participate in the Gandalf project. A lot has happened since we started until now that we are just about to finish the project. I remember our first meeting in April in Paris. Although troubleshooting was not even going to be part of the project, in that meeting we started to consider it. And now that we are at the end, troubleshooting has grown to be a third part of the Gandalf project. Thanks to all members of the Gandalf team for creating such a great working environment. Specially, thanks to Zwi Altman, for being such a good manager, so exhaustive in his work (he even followed every single algorithm that was developed for the project!) and open-minded to accept any proposal anyone in the group may have. Thanks to those who have worked with me in the troubleshooting part of the project: Lars Moltsen, Pedro Lázaro, Herve Dubreil, Beatriz Solana, Jordi Triola, Rana Khanafer and Sasikanth Munagala. I am looking forward to our next meetings in Berlin and Paris.

Finally, the circle has closed. The start of this thesis was diagnosis in satellite ground stations. Now that the Gandalf project has almost finished, we have just started a new ESA project for automatic diagnosis in satellite communication equipment. Thanks to Lars Moltsen for inviting me to participate in it and for believing in me.

Thanks to my two supervisors, Luis Díez and Volker Wille, specially for their help in making this document readable (what a disaster the first draft was!). Thanks to Finn Jensen and the people at his department for their warm welcome and support during the two weeks I spent in Aalborg. Thanks to José María Hermoso and José Antonio Fernández, who worked in troubleshooting with me for their final master thesis.

Thanks to all the people who have been with me all these years, fulfilling other areas of my life so necessary to make me focus on this thesis. Thanks to my parents and my sister for being always by my side, being content with the little time that my busy life left for being with them. I owe the numbers to my father and the letters to my mother. Thanks to Pedro for following me in every new restless challenge I undertake and for bearing the constant "I cannot do anything because of my thesis". Thanks to my close friends Pedro, Javi and Ale for being an essential component of my spare time and for being always there when I need them. Thanks to Álvaro for being with me so many years, despite the different countries that we have always had in between. Thanks to Marco, Javier, Juan, Kris and Lionel for making me miss living in Germany. Thanks to my friends from the painting class who helped me to forget about engineering for a few hours every week. Thanks to my colleagues in the Ingeniería de Comunicaciones Department, for creating such a nice working environment, specially to those who go down for coffee-time without having coffee and to those that have made me enjoy having lunch at university. Thanks to my students, who give sense to my work at university. And, finally, thanks to all the teachers that I have had along my life because this thesis is also due to them.

This thesis have been partially supported by:

- Nokia under a cooperation agreement (reference 8.07/59.1628) with the Ingeniería de Comunicaciones Group of the University of Malaga
- the Spanish Ministry of Science and Technology under project TIC2003-07827
- the Spanish Ministry of Industry, Tourism and Commerce under PROFIT FIT-330210-2005-42

Contents

Abstract

The mobile telecommunication industry has experienced significant changes in the recent past and it will continue to do so for the foreseeable future. The introduction of GSM based mobile communications in the early 90s has revolutionised the way people communicate. Prior to the launch of these systems, the majority of voice traffic was carried over plain old telephone systems and phones were associated to places rather than to people. At the same time, data communication has also undergone major changes. Due to applications such as e-mail and other internet based services, our communication behaviour is completely different from what it was less than 15 years ago.

The introduction of new features, services and technologies is still shaping and changing the communication industry. For example, the wide availability of WLAN-based broadband access and the launch of telephony based on WLAN is likely to revolutionise our communication pattern again. This change is potentially posing a real threat to the mobile communication business. Therefore, cellular network operators have to find ways to reduce the cost of their services and improve their quality to counter the threat posed by the emerging technologies.

In a mature cellular network that has undergone most of its site roll-out, the major cost is associated to the operation of the network. As the network consists of a high number of pieces of equipment that are distributed across the entire country, maintaining and operating this large and technically complicated system is a difficult task that requires operator personnel around the clock in several regional offices. For example, GSM networks in Europe may consist of about 10000 sites covering the entire country. Due to the large size of the networks, it is common that some of the deployed pieces of equipment do not work as planned. The consequence of such problems is poor end-user service. As in most countries several operators are competing for subscribers, it is imperative to rectify such occurrences because otherwise users will naturally switch to competing network operators. Hence, fault management, also called troubleshooting (TS), is a key aspect of the operation of a cellular system in a competitive environment. As the Radio Access Network (RAN) of cellular systems is by far the biggest part of the network, most of the TS activities are focused on this area.

TS comprises the isolation of faulty cells (fault detection), the identification of the fault causes (diagnosis) and the proposal and deployment of healing actions (solution deployment). Currently, in most cellular networks TS is a manual process, accomplished by RAN experts. Their task is to resolve problems in the network that have been identified by other employees or by automated checking routines. During the TS procedure, several applications and databases have to be queried to analyze performance indicators, cell configuration and alarms of the cells.

Amongst troubleshooting tasks, diagnosis of the cause of faults is the most complex and time-consuming one. Surprisingly, very few references can be found on automatic diagnosis in the RAN of cellular networks. Thus, the aim of this thesis is to study how to automate diagnosis for the RAN of cellular networks.

The first part of the thesis consists of a survey on how troubleshooting is currently performed in existing cellular networks. This part is considered to be essential for the rest of the thesis because the troubleshooting procedure is not documented in the existing literature, despite the fact that it is one of the main activities of operators of cellular networks.

Different techniques have been proposed for automatic diagnosis in other application domains. Amongst them, those based on Bayesian Networks (BNs) provide a modeling approach suitable to cater for the uncertainty inherent in human reasoning. This, as well as other benefits of BNs, together with the complexity of the problem under study, has focused this thesis on the study of automated diagnosis based on BNs.

In this thesis, two components of the diagnosis system have been distinguished: the model and the

inference method. The model represents the knowledge on how the identification of fault causes is carried out. The elements of the model are causes, symptoms and related parameters (named *conditions*). The inference method is the algorithm that identifies the cause of the problems based on the value of the symptoms and the conditions.

The model is determined by random variables, which represent the causes, symptoms and conditions, as well as probability functions, which model the relationship among these variables. Two types of diagnosis systems have been proposed in this thesis, depending on how symptoms have been modelled: either as continuous or discrete random variables. Firstly, the Bayesian Classifier (BC) models symptoms as continuous variables. In this case, the main difficulty in model definition lies in the specification of the conditional probability density function (pdfs) of the symptoms given the causes. In this thesis, it has been proposed to approximate those pdfs by beta pdfs.

In the second type of proposed systems, discrete BNs, the symptoms are modelled as discrete random variables. In this case, a technique traditionally used in order to simplify model construction is to assume a given network structure. That is, certain independence relationships among the variables are assumed. In that way, the problem of BN definition is simplified to the specification of the nodes and probability tables of a network with a given structure. In this thesis, different network structures have been selected, which take into account the fact that the simplicity in creating and using the model is a key issue in the cellular network domain.

Once the structure of the BN has been fixed, the parameters of the diagnosis model have to be defined, which can be done according to two solutions. Firstly, the model may be defined by diagnosis experts. Accordingly, in this thesis, a method to convert the specifications provided by experts in natural language into suitable diagnosis models has been described. Alternatively, the model can be constructed from training/example cases. Consequently, algorithms to learn the parameters of the model (thresholds for discretized variables and probabilities for the BN) have also been proposed in this thesis.

Contrary to other application domains, such as medical diagnosis, for which repositories of machine learning databases have been created and maintained, in cellular systems there are no databases of classified cases. This is the reason why in most cases the model has to be based on knowledge, i.e. experts define the model parameters. When the number of parameters to be set is large, which is normally the case, inaccuracy in setting these parameters is unavoidable due to multiple reasons. Hence, in this thesis two methods to minimise diagnosis error of models due to inaccurate parameters have been proposed.

Although the methods for automatic diagnosis proposed in this thesis are valid for any cellular network (2G, 3G or even future networks), they have been applied to diagnosis in GSM (2G). Thus, a secondary objective of this thesis has been to propose a model for automatic diagnosis in GSM. With this aim in mind, the main fault causes in GSM, their symptoms and conditions have been identified and interrelated.

Once the diagnosis systems are designed, they have to be evaluated. Firstly, in order to overcome the lack of cases from live networks, an algorithm has been developed to simulate training and test cases. Secondly, a methodology has been proposed to assess the performance and to compare the different diagnosis systems presented in this thesis. Two aspects have been considered: the performance, measured by different figures of merit, and the sensitivity of the results to imprecision in model parameters. Finally, the diagnosis systems presented in this thesis have been evaluated according to the previously described procedures.

Resumen

La industria de comunicaciones móviles ha experimentado cambios significativos en los últimos años y, previsiblemente, en el futuro continuará evolucionando con rapidez. La introducción de la telefonía basada en tecnología GSM a principios de los 90 revolucionó el modo en que la gente se comunicaba. Antes del lanzamiento de dicho servicio, la mayoría del tráfico de voz se transportaba por la red telefónica básica y los teléfonos estaban asociados con lugares en lugar de con personas. Al mismo tiempo, la comunicación de datos también ha sufrido grandes cambios. Debido a aplicaciones, como el e-mail y otros servicios basados en internet, las comunicaciones dentro de la sociedad son muy diferentes de lo que eran hace menos de 15 años.

La introducción de nuevas funcionalidades, servicios y tecnologías está aún moldeando las telecomunicaciones. Por ejemplo, es probable que la extensa disponibilidad de acceso inalámbrico de banda ancha y el lanzamiento de la telefonía basada en WLAN revolucione otra vez el mundo de las comunicaciones. Por tanto, los operadores de redes celulares deben encontrar medios para reducir los costes de sus servicios y mejorar la calidad para contrarrestar la amenaza de las tecnologías emergentes.

En una red celular madura que ha experimentado la mayor parte de su despliegue, el mayor coste está asociado a la operación de la red. Como la red consta de numerosos equipos distribuidos a lo largo de todo el país, mantener y operar este sistema complejo es una tarea difícil que requiere personal de forma permanente en muchas oficinas regionales. Por ejemplo, las redes GSM en un país europeo de tamaño medio pueden consistir en unos 10000 emplazamientos. Debido al gran tamaño de las redes, es frecuente que algún equipo no funcione como estaba planeado. Esos problemas tiene como consecuencia la prestación de un servicio deficiente a los usuarios finales. Como, en la mayoría de los los países, diversos operadores compiten por los clientes, es fundamental solucionar dichos problemas inmediatamente porque de otra forma los usuarios se cambiarán a redes de la competencia. Por tanto, la gestión de fallos, también llamada resolución de problemas (*troubleshooting*, TS), es un aspecto clave de la operación de un sistema celular en un entorno competitivo. Como la Red de Acceso Radio (*Radio Access Network*, RAN) de los sistemas celulares es la parte más importante de la red, la mayor parte de las actividades de TS se concentran en este área.

El TS engloba el aislamiento de las celdas con fallos (detección de fallos), la identificación de las causas de esos fallos (diagnosis) y la propuesta y realización de acciones correctivas (recuperación de fallos). Actualmente, en la mayoría de las redes celulares, el TS es un proceso manual, llevado a cabo por expertos en la RAN. Su tarea consiste en resolver los problemas de la red que han sido identificados previamente por otro personal de la empresa o por rutinas automáticas de chequeo. Durante el procedimiento de TS, se debe consultar múltiples aplicaciones y bases de datos para analizar los indicadores de funcionamiento, la configuración y las alarmas de las celdas.

Entre las tareas de gestión de fallos, la diagnosis es la más compleja y la que requiere más tiempo. Sorprendentemente, existen muy pocas referencias bibliográficas sobre la diagnosis automática en la RAN de redes celulares. Por eso, el objetivo de esta tesis ha sido estudiar como automatizar la diagnosis en la RAN de redes celulares.

La primera parte de la tesis ha consistido en una investigación sobre cómo se lleva a cabo la gestión de fallos en las redes celulares actuales. Esta parte se ha considerado esencial para el resto de la tesis, ya que el procedimiento de TS no está documentado en la literatura existente, a pesar de ser una las actividades principales de los operadores de redes celulares.

En otros dominios de aplicación se han propuesto diversas técnicas para la diagnosis automática. Entre ellas, las basadas en Redes Bayesianas (*Bayesian Networks*, BN) proporcionan una aproximación al modelado adecuada para tratar la incertidumbre inherente al razonamiento humano. Esto, además de otras ventajas de las BNs, junto con la complejidad del problema bajo estudio, han centrado esta tesis

en el estudio de sistemas automáticos de diagnosis basados en BNs.

En esta tesis, se han distinguido dos componentes del sistema de diagnosis: el modelo y el método de inferencia. El modelo representa el conocimiento de cómo se lleva a cabo la identificación de la causa de los fallos. Los elementos del modelo son causas, síntomas y parámetros relacionados (llamados *condiciones*). El método de inferencia es el algoritmo que identifica la causa de los problemas en función del valor de los síntomas y las condiciones.

El modelo queda determinado por variables aleatorias, que representan las causas, síntomas y condiciones, y por funciones de probabilidad, que modelan las relaciones entre las variables. En esta tesis se han propuesto dos tipos de sistemas de diagnosis, dependiendo de cómo se han modelado los síntomas: como variables aleatorias continuas o discretas. En primer lugar, el Clasificador Bayesiano modela los síntomas como variables continuas. En este caso, la principal dificultad en la definición del modelo radica en la especificación de las funciones densidad de probabilidades (fdps) de los síntomas dadas las causas. En esta tesis se ha propuesto aproximar esas fdps por funciones beta.

En el segundo tipo de sistema propuesto, BNs discretas, los síntomas se modelan como variables aleatorias discretas. Uno de los principales inconvenientes de las BNs es la dificultad para definir modelos complejos. Por eso, una técnica usada con frecuencia para simplificar la construcción del modelo es asumir una determinada estructura de red. Es decir, se suponen ciertas relaciones de independencia entre las variables. De esta forma, el problema de definir la BN se simplifica a la especificación de los nodos y tablas de probabilidad de una red con una estructura dada. En esta tesis, se han seleccionado diferentes estructuras, que tienen en cuenta que la sencillez en la creación y uso del sistema de diagnosis es un factor clave en la gestión de redes celulares.

Una vez que se ha fijado la estructura de la BN, se deben definir los parámetros del modelo de diagnosis, lo cual puede hacerse de dos formas. En primer lugar, el modelo puede ser definido por expertos en diagnosis. De acuerdo a esto, en esta tesis se ha descrito un método para convertir las especificaciones proporcionadas por expertos en lenguaje natural en modelos de diagnosis adecuados. Como alternativa, el modelo puede aprenderse a partir de casos de entrenamiento. Consecuentemente, se han propuesto también en esta tesis algoritmos para aprender a partir de datos los parámetros del modelo (umbrales para variables discretizadas y probabilidades para la BN).

Al contrario que en otros dominios de aplicación, como en el diagnóstico médico, en el que existen numerosas bases de datos para aprendizaje, en sistemas celulares no se dispone de dichas bases de datos de casos clasificados. Por esa razón, en muchas ocasiones, el modelo debe estar basado en el conocimiento, es decir, los expertos deben definir sus parámetros. Cuando el número de parámetros es demasiado grande, lo cual suele ocurrir con frecuencia, la imprecisión en la definición de los parámetros es inevitable debido a múltiples motivos. Por eso, en esta tesis se han propuesto dos métodos para disminuir los errores en el diagnóstico de modelos sujetos a parámetros imprecisos.

Aunque los métodos propuestos en esta tesis son válidos para cualquier red celular (2G, 3G o incluso futuras redes), se han aplicado a la diagnosis en GSM (2G). Por tanto, un objetivo secundario de esta tesis ha sido proponer un modelo para la diagnosis automática en GSM. Con este propósito, se han identificado e interrelacionado las principales causas de fallos en GSM, sus síntomas y condiciones.

Una vez que se han diseñado, los sistemas de diagnosis deben ser evaluados. En primer lugar, para suplir la falta de casos de redes reales, se ha desarrollado un algoritmo para simular casos de prueba y de entrenamiento. En segundo lugar, se ha propuesto una metodología para valorar el funcionamiento y comparar los distintos sistemas de diagnosis presentados en esta tesis. Se han considerado dos aspectos: el funcionamiento, medido mediante diferentes figuras de mérito, y la sensibilidad de los resultados a imprecisiones en los parámetros del modelo. Finalmente, los sistemas de diagnosis presentados en esta tesis se han evaluado de acuerdo a los procedimientos anteriormente descritos.

Acronyms

3G: Third Generation

AGCH: Access Grant Channel

AuC : Authentication Center

BA: BCCH Allocation

BC: Bayesian Classifier

BB FH: Baseband Frequency Hopping

BCC: Base Station Colour Code

BCCH: Broadcast Control Channel

BCF: Base station Control Function

BDF: Beta Distribution Function

BER: Bit Error Rate

BMAP: Beta Maximum a Posteriori

BN: Bayesian Network

BOG: Back Office group

BSC: Base Station Controller

BSS: Base Station Subsystem

BTS: Base Transceiver Station

CBM: Central Bayes Model

CDF: Cumulative Distribution Function

CF: Certainty Factor

CFG: Customer Faults group

CM: Communication Management sublayer

CN: Core Network

COTS: Commercial Off-The-Shelf

CSR: Call Success Rate

DAG: Directed Acyclic Graph

DAR: Diagnosis and Recovery tool

DCR: Dropped Call Rate

DL: Downlink

D-S: Dempster-Shafer

DTX: Discontinuous Transmission

EDGE: Enhanced Data Rates for Global Evolution

EIR: Equipment Identity Register

EMD: Entropy Minimization Discretization

EPS: Equivalent Prior Sample

ESA: European Space Agency

EXP: Probability definition based on experience

FACCH: Fast Associated Control Channel

FCCH: Frequency Correction Channel

FCPN: Fuzzy Causal Probabilistic Network

FD: Fault Detection

FDMA: Frequency Division Multiplex Access

fdp: función densidad de probabilidad

FER: Frame Erasure Rate

FH: Frequency Hopping

FOG: Front Office group

gaus: Gaussian belief mapping function

GERAN: GSM/EDGE Radio Access Network

GPRS: General Packet Radio Service

GSM: Groupe Spécial Mobile, Global System for Mobile Communication

HLR: Home Location Register

HO: Handover

HSN: Hopping Sequence Number

HW: Hardware

ICI: Independence of Causal Influence

ID: Identifier

IMEI: International Mobile Equipment Identity

IMSI: International Mobile Subscriber Identity

ISDN: Integrated Services Digital Network

ISO: International Organization for Standardization

ITU: International Telecommunications Union

KA: Knowledge Acquisition

KAT: Knowledge Acquisition Tool

KPI: Key Performance Indicator

LAPD: Link Access Protocol on the D channel

LAPDm: Link Access Protocol on the Dm channel

lrn: learnt parameters

MAIO: Mobile Allocation Index Offset

MAP: Maximum a Posteriori

MDL: Minimum Description Length

MEST: M-estimate

MHA: Masthead Amplifier

MLE: Maximum Likelihood Estimation

MM: Mobility Management sublayer

MS: Mobile Station

MSC: Mobile Switching Center

MOS: Mean Opinion Score

MPD: Minutes Per Dropped Calls

MUI: Multiple Uniform Intervals

net: network cases

NMS: Network Management System

NSS: Network and Switching Subsystem

OMC: Operations and Maintenance Center

OSS: Operation Subsystem

PC: Power Control

PCH: Paging Channel

PCM: Pulse Code Modulation

pdf: probability density function

PDF: Probability Density Function

PLMN: Public Land Mobile Networks

PSTN: Public Switched Telephone Network

PTA: Parque Tecnológico de Andalucía

QoS: Quality of Service

RACH: Random Access Channel

RAN: Radio Access Network

rect: Rectangular mapping function

ref: reference parameters

RF FH: Radiofrequency Frequency Hopping

ROC: Receiver operating characteristic

RR: Radio Resources Management sublayer

RRM: Radio Resource Management

RX: Receiver

SACCH: Slow Associated Control Channel

SBM: Simple Bayes Model

SBN: Smooth Bayesian Network

SCH: Synchronization channel

SDCCH: Stand alone Dedicated Control Channel

SEMD: Selective Entropy Minimization Discretization

SID: Silence Descriptor

sim: simulated cases

SIM: Subscriber Identity Module

SOM: Self-Organizing Map

Specs: GSM specifications

std: standard deviation

TA: Timing Advance

TCH: Traffic Channel

TDMA: Time Division Multiplex Access

TEXP: Discretization based on experience

trap: Trapezoidal belief mapping function

TRAU: Transcoder and Rate Adaptation Unit

TRX: Radio transmitter and receiver (transceiver)

TS: Troubleshooting

TSG: Technical Support group

TST: Troubleshooting Tool

TT: Trouble ticket

TX: Transmitter

UL: Uplink

UMTS: Universal Mobile Telecommunications Service

VLR: Visitor Location Register

WCDMA: Wideband Code Division Multiple Access

WLAN: Wireless Local Area Network

Chapter 1

Introduction

The aim of this chapter is to explain the purpose of this thesis and its motivation, to describe its objectives and to present the organization of this document.

1.1 Motivation

There is no doubt that during the last decade mobile communications have played an increasingly important role in telecommunications business, and it will continue to do so in the years to come. In most developed countries, the number of mobile telephones has exponentially growth over the last 10 years, in many countries even exceeding the number of inhabitants. Even for domestic use, the predominance of mobile over fix telephony is already a reality. For example, in Spain at the beginning of 2006 the penetration of mobile telephony at homes was 84.3%, whereas 83.5% of the homes had a traditional fix line [31].

The mobile communication industry is currently evolving from Second Generation (2G) towards Third Generation (3G). In the process, more functionalities and new technologies are gradually being added to the existing 2G networks. The first 2G cellular networks, which were introduced in the early 1990s, were based on Global System for Mobile communications (GSM) technology. The evolution began with an upgrade of the GSM network to 2.5G by introducing General Packet Radio Service (GPRS) technology. GPRS provides GSM with a packet data air interface and an IP-based core network. Enhanced Data Rates for Global Evolution (EDGE), which is based on the introduction of a higher speed modulation and improved coding scheme, was a further evolutionary step of GSM packet data. EDGE can handle about three times more data subscribers than GPRS, or triple the data rate for one end-user. GERAN (GSM/EDGE Radio Access Network) is the radio access technology based on GSM/EDGE. In the last years, 3G networks, called Universal Mobile Telecommunications Service (UMTS) networks in Europe, have started to be deployed throughout the world. In the near future, thanks to 3G, mobile internet-services are expected to be available “anywhere and anytime”. Users will surf the Web, check the email, download files or have real time videoconference, in a shopping mall, the airport, the city center or their offices.

Thus, the current scenario comprises a complex set of interrelated and rapidly growing wireless networks, applications which require increasing bandwidth and users who demand high quality of service at low cost, but with a limited spectrum. In a few years, the highly complex and heterogeneous RAN will comprise different technologies, such as GSM, UMTS and WLAN[1]. As a result, the management of the RAN will be a tough challenge that the operators will have to tackle.

In addition, new technologies are foreseen to revolutionise the industry of telecommunications. For example, the increasing expansion of IP telephony, the wide availability of WLAN based broadband access or the launch of telephony based on WLAN (e.g. Skype [6]) is potentially posing a real threat to the business of mobile operators. Therefore cellular network operators have to find ways to reduce the cost of their services and improve their quality to counter the threat posed by the emerging technology.

In a mature cellular network that has undergone most of its site roll-out, the major cost is associated to the operation of the network. As the network consists of a high number of pieces of equipment that are distributed across the entire country, maintaining and operating this large and technically complicated system is a difficult task that requires operator personnel around the clock in several regional offices. For example, GSM networks in Europe may consist of about 10.000 sites covering the entire country. Due to the large size of the networks, it is common that some of the deployed equipment does not work as planned. The consequence of such problems is poor end-user service. As, in most countries, several operators are competing for subscribers, it is imperative to rectify such occurrences because otherwise users will be dissatisfied with the service and thus will likely switch to competing network operators. Hence, fault management, also called troubleshooting (TS), is a key aspect of operating a cellular system in a competitive environment. As the Radio Access Network[2] (RAN) of cellular systems is by far the biggest part of the network, most of the TS activities are focused in this area.

TS comprises the isolation of faulty cells (fault detection), the identification of malfunctions (diagnosis) and the proposal and deployment of healing actions (solution deployment/fault recovery). Currently, in most cellular networks, TS is a manual process, accomplished by RAN experts. These engineers are applying a series of customized checking routine on a daily basis in order to identify the cause of the problem. During the procedure, several applications and databases have to be queried to analyze performance indicators, cell configurations and alarms. The speed in identifying faults is dependent on the level of expertise of the troubleshooter, the type of information available and the quality of the tools displaying relevant pieces of information. This means that, in addition to a good understanding of the possible causes of the problems, a very good understanding of the tools available to access the sources of information is also required.

In this scenario, the benefits of automating TS are numerous. With the help of an automated TS tool, the time required to identify the reason for a fault causing a problem is greatly reduced.

[1]WLAN: Wireless Local Area Network

[2]The Radio Access Network (RAN) is the part of the network in charge of providing access to the users and connecting them to the core network. The wireless access is one of the key characteristics of mobile communication networks, which identifies them and differentiates them from traditional wired networks.

This means that network performance is enhanced as the downtime and the time with reduced quality of service (QoS) is significantly limited. In addition, by automating the TS process, fewer personnel and, thus, fewer operational costs are necessary to maintain a network of a given size. The TS process is de-skilled as the majority of problems can be rectified with the help of the automated TS tool. Then, the knowledge of highly experienced staff, which is released from the TS work, can be utilized for other aspects of network optimization, thereby further increasing network performance. One additional benefit is that the knowledge in TS can be stored in the TS tool, therefore not being dependent on the staff working for the company at any time. In conclusion, the gains achieved thanks to automated TS for an operator are significant as fewer personnel with a lower skill level can solve more network problems in less time.

The first steps in automation of troubleshooting in the RAN of cellular networks have been focused on performance visualization and on fault detection (FD) [124, 47, 132, 123, 104, 133, 48]. On the one hand, thanks to methods to achieve efficient visualization of the network performance, FD and diagnosis are carried out more easily. On the other hand, several methods have been proposed for FD, which consist in building models for the normal behavior of the system. The deviations of the available measurement variables from the normal behavior can then be detected with some type of abnormality detector.

Amongst troubleshooting tasks, diagnosis of faults is the most complex and time-consuming one. However, very few references can be found on automatic diagnosis in the RAN of cellular networks. This is the reason why in this thesis other application domains where diagnosis is also required have been surveyed. Based on that survey, the selected technique in this thesis has been Bayesian Networks (BNs). BNs, also called belief probabilistic networks, have been proposed by many authors as the modelling technique for the development of automatic diagnosis systems. Some application domains are the diagnosis of diseases in medicine [34, 146, 99, 199, 35, 154, 147, 144], the troubleshooting of printer failures [96, 51, 100, 94, 98, 101, 174, 111, 175, 173], the diagnosis of faults in satellite communication systems [107, 131, 57] and fault identification in the core network (CN) part of communication networks [115, 182, 190, 73]. The latter is the closest field to the RAN, although it has very important differences. On the one hand, in CNs, it is very important to represent the dependencies among communication system entities because, due to the highly interconnected architecture, a failure in an entity may have a large impact on other system entities. On the other hand, in general, in CNs the only type of symptoms are the alarms related to the network entities. In the RAN part of cellular systems, this interconnection is not as strong and alarms are not the only symptoms of malfunction, due to two main reasons. Firstly, most faults are not related to a physical component, but to a poor network planning or an incorrect parameter setting. Thus, modelling the interactions among entities is not a requirement like in other communication networks because, in most cases, faults are not related to pieces of equipment which are physically connected. Secondly, although alarms play a very important role in identifying faults in specific pieces of equipment, they do not provide conclusive information in order to isolate configuration problems. Furthermore, the fact that performance indicators in the RAN domain are continuous, instead of binary like alarms, generates a new difficulty in the modelling, which is nonexistent in the CN domain.

Research studies in automation of diagnosis in the RAN of cellular networks have been focused on alarm correlation [193, 84, 192, 88]. In current cellular networks, most systems are semi-intelligent and generate alarm messages when errors occur. The abstraction level of these alarm messages is normally very low, leading to a high number of alarms for any single cause. For example, when a link fails, up to 100 or more alarms are generated and passed to the Network Management System. Those alarms should be converted into a minimum number of alarms which clearly pinpoint to the breakdown of a link. Alarm correlation [109, 202] consists in the conceptual interpretation of multiple alarms, so that new meanings are assigned to the original alarms. It is a process that involves different tasks: reduction of multiple occurrences of an alarm into a single alarm, inhibition of low priority alarms in the presence of higher priority alarms, substitution of a specific set of correlated alarms by a new one, etc. This way, alarm correlation systems are required to filter and condense the incoming alarms to meaningful high level alarms in order to avoid overloading the operators. Although alarm correlation can be considered a first step in the diagnosis of faults, alarms do not provide enough information to identify the cause of problems, especially if the possible causes are not only faults in pieces of equipment. Even after alarm correlation, the number of triggered alarms for a single cause is normally very high. In addition, the same alarms may be triggered by different causes. Due to these reasons, in order to achieve a conclusive diagnosis, it is important that not only alarms are taken into account, but also network performance indicators. Because of the lack of studies related to this last aspect, this thesis is focused on automatic diagnosis in the RAN of cellular networks based on performance indicators.

1.2 Preliminaries

In the year 2000, Nokia Networks opened a Research Center for Mobile Communications at the *Parque Tecnológico de Andalucía* (PTA) in Málaga. This center was created in the framework of a cooperation agreement between Nokia and the University of Málaga (in particular with the *Grupo de Ingeniería de Comunicaciones*). The staff was composed of experienced Nokia personnel, as well as more than 50 new employees and lecturers from University of Malaga.

One of the projects that was started in the new Nokia center, in 2001, was a Troubleshooting project. Its aim was to design an automatic troubleshooting tool for the RAN of cellular networks. Nokia considered it essential to open a line of research in this area because of the reasons explained in Section 1.1. Building an automatic troubleshooting tool was not trivial at all as it required knowledge from very diverse fields. To begin with, it was necessary to have the theoretical knowledge to select the most appropriate artificial intelligence method for the area under study. Then, there was a need to fully understand the area to which this method could be applied, i.e. the troubleshooting in the RAN of a cellular network. Finally, the skill of turning this knowledge into a usable tool that interfaces with the other hardware and software already in use by the operator was essential. After several fruitless attempts by first Nokia alone, and then Nokia and the University of Malaga, it became clear that only the close co-operation between a network operator, a research unit and an equipment manufacturer could provide the three above-mentioned skills. Firstly, an equipment vendor like Nokia, can provide alarms (i.e. "when a value is below/above a certain level") and performance indicators, but only TS engineers from network operators know about the actual meaning (or lack of meaning) of such alarms and performance indicators in actual networks. This knowledge can not be simulated or predicted, but it has to be acquired via years of hands-on experience. However, although the TS engineers have the TS knowledge, they generally lack the theoretical bases on automatic TS systems (e.g. which artificial intelligence technique to use, how this work, how to create the TS system, etc). Secondly, engineers from manufacturers of network equipment are not only experts in equipment, but they also have a general knowledge on mobile communications. Finally, researchers from university provide their expertise in theory of mobile communications and artificial intelligence.

The troubleshooting group was composed of staff in Spain, UK and Denmark. Moreover, the group worked with operators in UK and Denmark. My responsibility was the research part of the project, and that constituted the starting point of this thesis. Later, I was not only involved in research, but also in the specification of prototypes for diagnosis and in meetings with cellular networks operators.

Operators would only be interested in a cooperation if the TS tool could demonstrate its ability in their networks. Furthermore, in order to define an accurate model, the prototype had to be constantly tested with operators. As the initiative was driven by Nokia, the aim was to provide a "platform" on which the operator knowledge could be first gathered and then executed. Thus, two prototype tools were implemented: a troubleshooting tool, which carried out the automatic diagnosis, and a knowledge acquisition tool, which obtained the knowledge from TS experts and built the TS model. In addition, a TS model for GERAN was created thanks to the cooperation with the network operators. Bayesian Networks were the technique used to deploy the automatic TS system. The TS model was based on knowledge only, that is, no previous examples were used to obtain the parameters of the model.

Unfortunately, due to major re-structuring, Nokia closed its entire center in Málaga in June 2003 and, consequently, the troubleshooting project was also stopped. However, as defining an automatic TS system was still an important pending issue in current and future cellular networks, and the objectives of my thesis were clear, my research continued on this topic.

Some months later, other projects came to support this thesis. Firstly, in 2003, the Spanish Ministry of Science and Technology funded a research project of the Communication Engineering Department named "Radio resources optimisation tools for mobile communication networks". One of the lines of this project was automatic diagnosis.

In addition, in 2004 a project called *"Gandalf: Monitoring and self-tuning of RRM parameters in a multi-system network"* [3, 2] was started by a consortium composed by France Telecom R&D, the University of Limerick, Telefónica R&D, Moltsen Intelligence Software and the University of Málaga. One of the work packages was automatic diagnosis in multi-system cellular networks. This project was started due to the continuing interest of cellular operators in this area of work. The project was assigned a CELTIC label [1] within the EUREKA European network [4]. As one of the results of this project, a commercial automatic troubleshooting tool, named Moltsen TheCure [5], has been deployed.

1.3 Objectives

The main aim of this thesis is the design of an automatic diagnosis system for the RAN segment of cellular networks. Fig.1.1 summarizes the scenario and the location of the specific objectives (squared yellow boxes) of this thesis. Firstly, a *case* is defined as a set composed of a fault cause in a cell and the value of its symptoms and related parameters. As observed in Fig.1.1, cases can be obtained from two sources: a live cellular network and a case simulator. At this point, one of the objectives of the thesis (5.a) is to define methods to simulate cases. Cases can be used for two different purposes: as training examples for model learning or as test cases for diagnosis. In the former, another aim of the thesis can be found (4.a): to design methods to learn the model parameters based on training cases. Alternatively, model parameters can be elicited by diagnosis expert. Hence, another objective (4.b) is to develop techniques to ease knowledge acquisition. Experts may build a specific model for diagnosis in GERAN. The definition of that model is defined in objective (2). Another objective (3.a) is the proposal of diverse techniques to model diagnosis. In Fig.1.1 different models (Model 1, Model 2) are represented, which differ in the modelling technique applied to build them. Once the model is defined, and their parameters are incorporated, it is used by the diagnosis system to assess the fault cause. At this point, another objective (3.b) of the thesis is to define methods for automatic diagnosis. The inputs to the diagnosis system are the symptoms and configuration parameters in the test cases. In order to evaluate the performance of the diagnosis system, its output, i.e. the diagnosed fault cause, is compared with the real cause. At this point, an objective (5.b) is the definition of methods to evaluate and compare diagnosis systems. In addition to the objectives shown in Fig.1.1, a prior objective is the analysis of the state-of-the-art (1).

In summary, the objectives of this thesis can be structured in:

1. **Problem approach.** At a first stage, a survey on how troubleshooting is currently performed in existing cellular networks should be carried out. This part is considered to be very important because the troubleshooting procedure is not documented in the existing literature. In addition, different techniques for automatic diagnosis, which take into account the inherent uncertainty in human reasoning, should be examined.

2. **Diagnosis in GERAN.** The techniques for automatic diagnosis proposed in this thesis should be valid not only for 2G networks, but also for 3G and future cellular networks. Nevertheless, a secondary objective of this thesis is to propose a system for automatic diagnosis in GERAN. With this aim in mind, the main fault causes in GERAN, their symptoms and other related parameters should be identified and interrelated.

3. **Modelling of diagnosis.** The main objective of the thesis is to develop methods (3.b) to automatically diagnose the fault cause in the RAN of any cellular network. Those methods use a diagnosis model, which represents the knowledge on how the identification

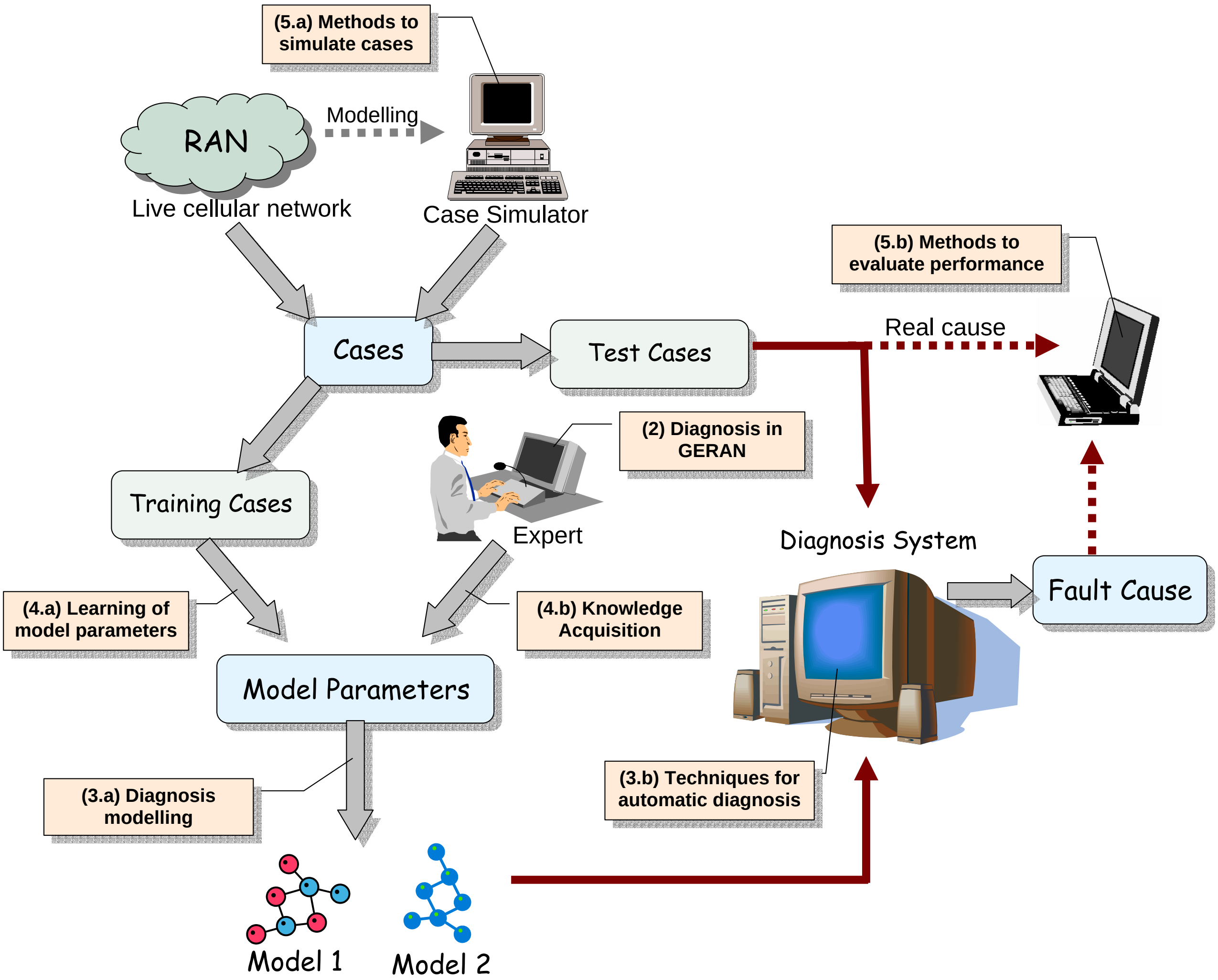


Figure 1.1: Objectives of the PhD

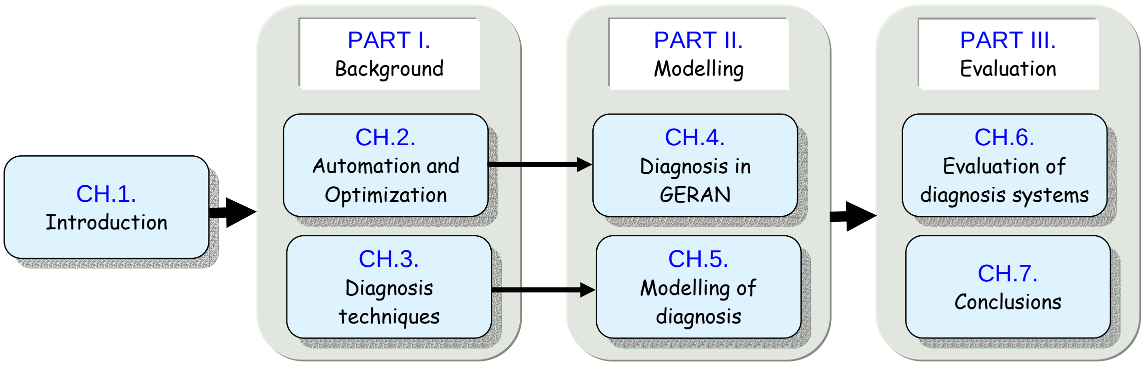


Figure 1.2: Organization of the thesis report

of faults should be done. The main elements of the model and its parameters should also be investigated at this stage (3.a).

4. **Model construction**. In order to define the parameters of the diagnosis model, two alternative solutions should be analyzed. On the one hand, the model may be defined by diagnosis experts. Hence, an objective (4.b) will be to study how to convert the specifications provided in natural language by those experts into diagnosis models. This is called knowledge acquisition. On the other hand, the model can be learnt from training examples. Thus, another target (4.a) will be to develop methods to automatically learn the parameters from those training cases.

5. **Model evaluation**. The last objective is to design techniques to evaluate diagnosis systems. Firstly (5.a), an algorithm should be developed to generate cases, that is to simulate the behavior of the network in the presence of faults (*Case simulator*). Secondly, a set of evaluation methods will be proposed (5.b). They will be used to assess the performance and to compare the different diagnosis systems presented in this thesis. Two different aspects will be considered: the performance measured by different figures of merit and the sensitivity of the results to imprecision in model parameters.

1.4 Guide for the reader

The organization of this thesis report (Fig.1.2) is directly related to the objectives presented above. The first chapter corresponds to this introduction to the thesis. The rest of the document is organized in three parts.

The first part includes the required background to follow the rest of the report. This part comprises a review of the state-of-the-art of techniques used and of the theoretical basis of the areas covered in this thesis. Due to the interdisciplinary nature of this work, this part has been divided into two clearly distinguishable chapters within the disciplines of mobile communications and artificial intelligence. Chapter 2 carries out a survey on automation and optimization and on how troubleshooting is performed in current cellular networks. In addition, a brief introduction to GSM is provided. Chapter 3 is focused on artificial intelligence techniques used in other disciplines for automatic diagnosis. Amongst them, the most appropriate technique for cellular networks is selected.

The second part of this document is devoted to the design of automatic diagnosis systems. It is also composed of two chapters. Chapter 4 studies diagnosis in GERAN. Hence, the main causes, symptoms and related parameters to be taken into account in diagnosis are analyzed in this chapter. Chapter 5 proposes models and techniques to automate diagnosis, valid not only for 2G networks, but also for 3G or future cellular networks.

Finally, the third part of the thesis is dedicated to the evaluation and comparison of the systems proposed in Part II. In Chapter 6 the methodology to obtain simulated cases is described. In addition, the design of the experiments, the methods for sensitivity analysis and the figures of merit will be also explained in this chapter. Results obtained with the systems proposed in Part II, following the previously explained evaluation methodology, are also presented in this chapter and in appendixes A-D. Chapter 7 summarizes the main conclusions of the research and proposes future lines of action.

This report also includes as appendix E a summary of the thesis in Spanish.

Part I

Background

Chapter 2

Automation and optimization in cellular networks

This chapter outlines the importance of automation and optimization in cellular networks. In the last few years, driven by the increasing complexity of networks, the attention of cellular network operators is turning to the study of methods to achieve more efficient network management. On the one hand, the multi-service capabilities of 3G networks will bring new challenges to network operation and maintenance. On the other hand, in the very near future, different radio access technologies, such as GSM, UMTS and IEEE 802.11 (WLAN) will coexist within the same network. It will not be feasible to carry out the operation of these multi-system networks with the manual procedures employed in current networks. Therefore, automation of network management is crucial to face the forthcoming changes in the mobile telecommunication industry.

In this thesis, a methodology for automatic troubleshooting of the radio access part of cellular networks is proposed. The described methods are valid for any network, such as 2G, 3G or WLAN networks if minor modifications are applied. The proposed methodology has been applied to build a diagnosis model for GSM/GPRS networks. The reason for choosing GSM is five-fold: 1) the existence of live networks from which to obtain data to build the models and test the designed systems, 2) well established processes for TS (although not well documented in the public domain), 3) stability of the software and HW used by operators 4) strong operator interest due to cost pressure, 5) GSM will be used for many years to come.

This chapter is divided into two parts. In the first part, the basics of GSM networks are summarised. Special attention is drawn to those aspects of GSM required to follow the rest of this thesis. The second part of the chapter is a survey on automation in mobile communication networks. Firstly, the importance of automation and optimisation in cellular networks is highlighted. Subsequently, it is described how troubleshooting is carried out in current cellular networks. Finally, the principles of automatic troubleshooting are presented and the state of the art is analyzed.

2.1 Preliminaries

2.1.1 Overview of the GSM system

The main characteristic of Public Land Mobile Networks (PLMN) is that they allow the user to move within their coverage area. The coverage area of a given operator is divided into geographical areas referred as *cells*, each of them covered by a base station, i.e. mobiles in this cell can connect to the fixed part of the network via that base station through a radio link. The user interface to the PLMN is the *mobile station* (MS).

Currently, the most deployed cellular network standard is by far GSM. Although the techniques proposed in this thesis are also valid for other cellular networks, such as UMTS, models have been built for GSM networks. Some basic characteristics and definitions related to GSM are the following:

- Good speech quality comparable to that of fixed networks.
- Support for new services and facilities.
- Authentication of users and equipment.
- Security in the calls.
- Interworking among equipment from different manufacturers.
- Interworking with the existing fixed networks (PSTN[1], ISDN[2], data networks, etc.).
- International *roaming*: with a unique number, clients can generate and receive calls from most countries in the world.
- *Location*: at any moment the network must know where each subscriber is situated, so that incoming calls can be routed.
- *Paging*: when there is an incoming call, the network must broadcast a signal to the MSs in the *location area*[3] where the MS destination of the call is situated.
- *Handover (HO)*: it is the switching of an on-going call[4] to a different channel or cell. The continuity of the call should be assured by the network.

There are numerous books and journals related to mobile communications and, in particular, to GSM. For an introduction, readers are referred to [142, 161, 162, 145, 90, 7]. The following sections are focused on those aspects of GSM that are more related to this thesis. In Section 2.1.2, the architecture of the GSM network will be described. In Section 2.1.3, signalling protocols in the radio interface will be summarized. In Section 2.1.4, the channel structures will be enumerated. Finally, in Section 2.1.5, network management will be introduced.

[1]PSTN: Public Switched Telephone Network

[2]ISDN: Integrated Services Digital Network

[3]Location area: area in which a MS can move without changing its location registry in the network. When there is an incoming call, a paging is sent to all MSs in the location area

[4]a MS can be in two states: *active*, during an on-going call, or *idle* when not in a call.

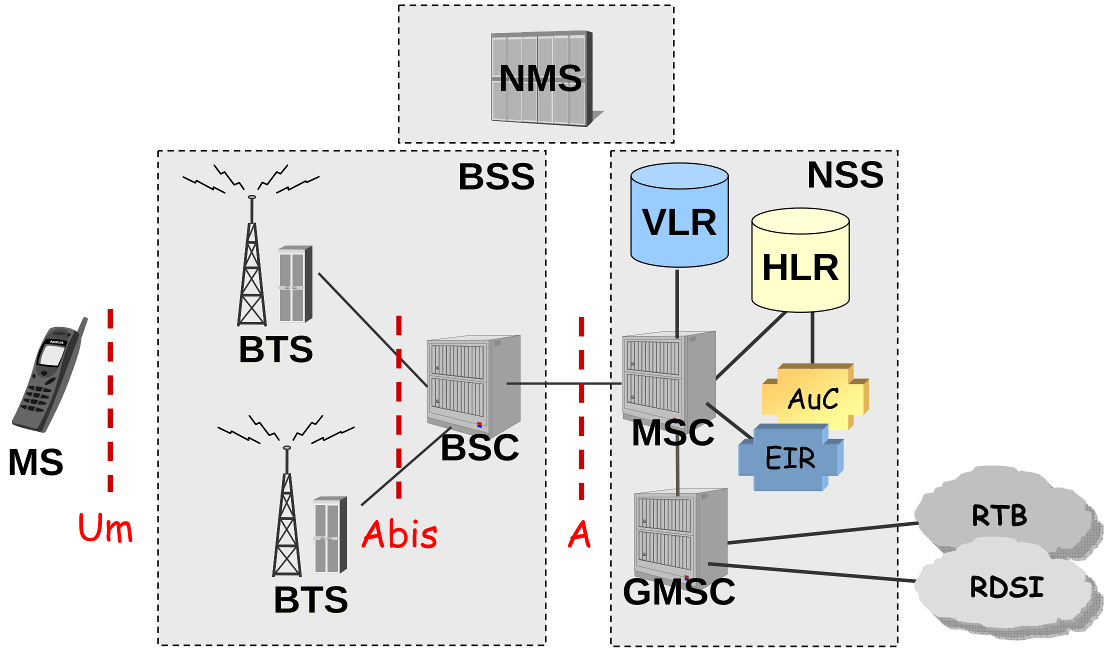


Figure 2.1: Architecture of a GSM network

2.1.2 Architecture of the GSM network

The GSM network is hierarchically organized in subsystems (Fig.2.1), which in turn are composed of functional entities. The *Base Station Subsystem* (BSS) controls the radio links with the MSs. The *Network and Switching Subsystem* (NSS) performs the routing of calls and handles the mobility management operations. The *Operations and Maintenance Center* (OMC), also named *Network Management System* (NMS) and *Operation Subsystem* (OSS) is used to oversee the proper operation and setup of the network. The interfaces between the subsystems and entities are denoted by letters. The interface between the MSs and the network is the *radio interface*, also referred as *air interface* or U_m *interface*. The BSS is linked to the NSS through the *A interface*.

The MS consists of the mobile equipment (the terminal) and a smart card called the *Subscriber Identity Module* (SIM). The SIM provides personal mobility, so that the user can have access to subscribed services irrespective of the specific terminal. The mobile equipment is uniquely identified by the *International Mobile Equipment Identity* (IMEI). The SIM card contains the *International Mobile Subscriber Identity* (IMSI) used to identify the subscriber to the system, a secret key for authentication, and other information.

The BSS is in charge of providing and managing transmission paths between the MSs and the NSS. It contains the *Base Transceiver Stations* (BTS) and the *Base Station Controller* (BSC), which communicate across the *A-bis interface*. There is a BTS in each cell, which contains all the radio transmission and reception equipment: radio transmitters and receivers (TRX), antennas, connection elements to the antennas, supplementary elements (towers, air conditioning, etc.). The BTS is also in charge of the signal processing. The BSC controls several BTSs and it handles the radio interface management, e.g. allocation and release of radio channels and handover management.

The NSS includes the main switching functions of GSM and the databases required for subscriber data and mobility management. The main function of the *Mobile Switching Center* (MSC) is call routing. In addition, it provides other functionalities, such as the connection to the fixed networks (e.g. PSTN or ISDN), registration, authentication, location updating and handovers. There are four different types of handover in GSM, which involve transferring a call between: channels (time slots) in the same cell, cells (BTS) under the control of the same BSC, cells under the control of different BSCs, but belonging to the same MSC and cells under the control of different MSCs. The first two types of handover, called internal handovers, involve only a single BSC. They are managed by the BSC without involving the MSC, except to notify it at the completion of the handover. The last two types of handover, called external handovers, are handled by the involved MSCs. The MSC works in conjunction with several functional entities. Thus, associated to each MSC there is a temporal database named *Visitor Location Register* (VLR), which contains dynamic administrative information for each mobile currently located in the geographical area controlled by that VLR, necessary for call control and provision of the subscribed services. The *Home Location Register* (HLR) contains all the administrative information of each subscriber registered in the corresponding GSM network, along with the current location of the mobile. The location of the mobile is typically in the form of the signalling address of the VLR associated with the mobile station. The other two registers are used for authentication and security purposes. The *Equipment Identity Register* (EIR) is a database that contains a list of all valid mobile equipment on the network, where each mobile is identified by its IMEI. The *Authentication Center* (AuC) is a database that stores a copy of the secret key saved in each subscriber's SIM card, which is used for authentication and encryption over the radio channel.

The NMS interacts with the BSS and NSS subsystem and it handles most aspects of the network management. The main functions of the NMS can be divided into three categories: fault management, performance management and configuration management. The purpose of the Fault Management is to ensure the smooth operation of the network and the rapid correction of any kind of problems that are detected. Fault management provides the network operator with information about the current status of alarm events and maintains a historical database of alarms. The purpose of the Configuration Management is to maintain up to date information about the operation and configuration status of the network elements and to enable efficient handling of these. In Performance Management, the NMS collects measurements from individual network elements and stores them in a database. On the basis of these data, the network operator can compare the actual performance of the network with the planned performance and detect both good and bad performance areas within the network.

A typical network in a country the size of Spain may be composed of about 25000 BTSs, 200 BSCs, 60 MSCs and 3-10 NMSs.

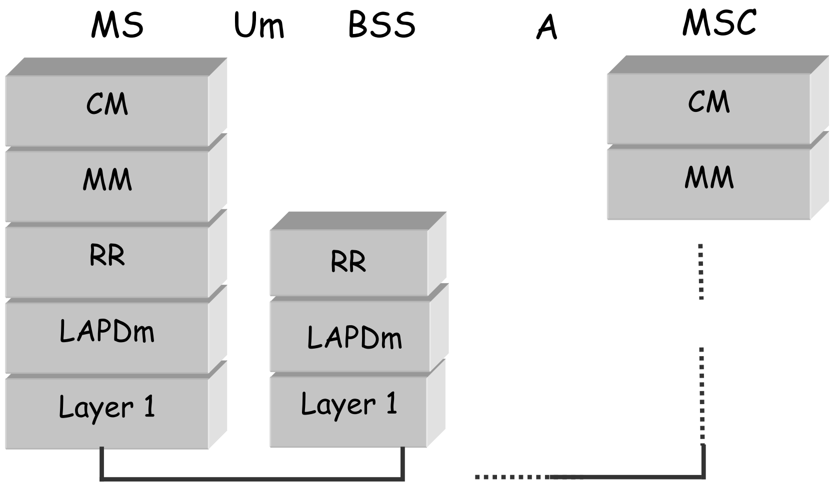


Figure 2.2: Signalling protocol structure in GSM

This thesis is focused on the *Radio Access Network* (RAN[5]), which guarantees the connection between the MSs and the core network. The RAN in GSM/GPRS, which is called GERAN, comprises the MSs and the BSS.

2.1.3 Signalling protocols in the radio interface

Fig.2.2 shows the signalling protocols in the GSM RAN. Layer 1 is the physical layer of the radio interface, which contains the functionalities required to transmit the information over the radio channels. Hence, the physical layer is organized in physical and logical channels (see Section 2.1.4). The physical layer is in charge of associating physical and logical channels, channel coding, information ciphering, cell selection in idle mode and supervision of the communication quality.

The aim of layer 2 is to establish a secure and reliable signalling link between the MS and the network. The protocol is called LAPDm, which is based on a modification of the LAPD protocol used in ISDN.

Layer 3 is in charge of signalling between the MS and the network and it is composed of three sublayers: RR, MM and CM. The *Radio Resources Management* (RR) sublayer controls the setup, maintenance, and termination of both radio channels and connections with the MSC. In addition, it handles the management of radio features such as power control, discontinuous transmission and timing advance (see Chapter 4). The execution and measurements required for handovers are also responsibility of the RR layer. The *Mobility Management* (MM) sublayer handles the functions that arise from the mobility of the subscriber, as well as the authentication and security aspects. The former implies the required procedures to update the location in the VLR and HLR. Security includes the authentication of users to prove that they are who they

[5]RAN: The component parts of the mobile radio network infrastructure that manage and facilitate communications between the mobiles and the core network

claim to be, identification of mobile equipment, ciphering and assignation of temporal identities. The *Communication Management* (CM) sublayer is responsible of procedures related to call control and set-up, supplementary service management, and short message service management.

2.1.4 Channel structures

The GSM radio interface combines Frequency Division Multiplex Access (FDMA) with Time Division Multiplex Access (TDMA). On the one hand, the frequency band is divided into 124 carrier frequencies spaced 200 kHz apart. One or more carrier frequencies are assigned to each BTS. The radio links are duplex, that is two different carriers are used for the uplink[6] and for the downlink[7]. Each pair of frequencies is denoted as a *radiochannel*. On the other hand, the time is divided into frames and each frame is divided into 8 *time slots* of about 0.577 ms. In this way, 8 physical channels are provided over each radiochannel. Therefore, a bidirectional *physical channel* between a BTS and a MS is defined by the radiochannel frequency and the time slot assigned to that communication.

A *logical channel* is an information stream dedicated to the transfer of a specific type of information over the radio interface [9]. Logical channels can be classified depending on their usage into two groups: *common channels* and *dedicated channels*. Common channels are those channels which transmit signalling information common to all mobiles in a cell. Amongst all the carriers transmitted in a cell, one of them works as beacon carrier and its time slot 0 is used to transmit the common channels. This carrier is denoted as *BCCH carrier*, because the BCCH logical channel goes onto that carrier. The most important common channels are:

- **FCCH (Frequency Correction Channel):** It carries information from the BSS for carrier synchronization.
- **SCH (Synchronization channel):** It carries information from the BSS for frame synchronization.
- **BCCH (Broadcast Control Channel):** It broadcast on the downlink information such as the BTS identity or the location area. This channel is continuously active because its signal strength is monitored by mobiles for handover purposes.
- **PCH (Paging Channel):** It is used by the network to call for a mobile when it has an incoming call.
- **RACH: (Random Access Channel):** it is used by the mobiles when they need to access the network.
- **AGCH (Access Grant Channel):** It is used by the system to assign a dedicated channel to a mobile.

[6] Uplink: transmission path from the MS to the BTS
[7] Downlink: transmission path from the BTS to the MS

Dedicated channels transmit information of an established connection between a MS and the network. The most common dedicated channels are:

- **TCH/F (Traffic Channel Full Rate) and TCH/H (Traffic Channel Half Rate):** They are the bidirectional traffic channels used to carry speech and data. There are two types of TCH: full rate and half rate, which are distinguished by the channel period.

- **SACCH (Slow Associated Control Channel):** It is the signalling channel associated to a traffic channel. It is used to transmit the required information to manage the radio resources, such as downlink level and quality measurements.

- **FACCH (Fast Associated Control Channel):** It is a signalling channel associated to a traffic channel. It is used to transmit information that cannot wait until the corresponding SACCH arrives, such as the handover information.

- **SDCCH (Stand alone Dedicated Control Channel):** This channel is used for call setup, location updating and transmission and reception of short messages.

Traffic channels are defined using a 26-frame multiframe (i.e. a group of 26 TDMA frames). Out of the 26 frames, 24 are used for traffic, one is used for the SACCH and one is currently unused.

2.1.5 Network management

Diagnosis is one of the main tasks in network management. Network management is responsible for the efficient operation and organization of telecommunication networks. ISO[8] together with ITU[9] standardized network management following the OSI Reference Model. Functional areas defined by the standard are:

- Fault management, which is responsible for detection, isolation and correction of network faults.

- Configuration management, which provides the operators with the means to define, control and monitor network elements in order to maintain a reliable communication network

- Accounting management, which deals with managing the billing and charging system, calculating the cost of network services

- Performance management, which handles the execution of performance measurements by monitoring and analysing the managed network elements and services

- Security management, which ensures that the information exchanged by the network is not corrupted.

[8]ISO: International Organization for Standardization
[9]ITU: International Telecommunications Union

Network management has always lagged behind network technology development due to the following factors [176]:

- Management is seen as a non-functional requirement of networking systems and hence is given lower research priority despite its importance for operators
- Management functions, especially configuration management, depend on the attributes and capabilities of the network entities being managed and are mostly vendor-specific
- Network management is by its nature a multi-vendor and multi-technology integration problem; this ensures that turn-key solutions are very difficult to develop.

Current networks, such as GSM, UMTS and WLAN still require significant manual configuration and management for deployment and operation. However, the rapid increase in complexity and size of the networks being managed have led to a widespread belief that current management models need to change to meet the challenges of future ubiquitous networking. Hence, recent trends in network management research include the concept of automatic or self-organising systems (e.g. see [30] for self-organisation in mobile networking), although these are not present in existing networks.

Amongst the functional areas of network management, in addition to fault management, this thesis is specially focused on network performance because of its relation with diagnosis based on performance indicators.

Network performance

In order to monitor network performance, three sources of information are normally considered [14]: customer complaints, field tests and statistics in the NMS.

Firstly, quality improvement should take place before customer complaints are numerous. Based only on customer complaints it is difficult to distinguish whether the problem is in the MS or in the network. Therefore, it is not efficient to optimize the whole network using only this source of information. However, with good cooperation between the customer service and the network optimization team within the company, it should be possible to take advantage of customer complaints to improve network quality. For example, sometimes customer complaints can help to find not known problems and to identify their geographical location.

Secondly, field tests are very resource and time consuming. They are normally restricted to specific areas and, in most cases, they only supply information about the downlink path. Therefore, optimization of the whole network is also not possible based only on drive tests, although they may help to locate problems in certain areas. Furthermore, field tests are useful to adjust propagation models and to pinpoint coverage holes. Nevertheless, a common problem within operators is that data from the field tests are used to inform managers about the performance of the network, but not to optimize it.

Therefore, statistics from the NMS are the main source of information used to determine network quality and to carry out troubleshooting or optimization. The most important events over a reporting measurement period (typically one hour) related to each cell are collected by the BSC to which the cell is connected. These events can belong to two types: *measurements*, also called *counters*, or *alarms*. Both counters and alarms are transmitted and stored in the NMS. Standard measurements are collected continuously, whereas special ones may only be switched on when required. Data in the NMS is usually stored for several days. The main drawback of this source of information is that statistically relevant volumes of traffic are required to provide reliable results. However, it has many advantages compared to the other sources mentioned earlier: it is a cost efficient way to monitor network quality, limited geographically location of problems is possible, it allows centralized data collection, it can be used to monitor trends, locating problems on a per-cell level is possible, etc.

Measurements related to the RAN can be classified into the following types:

- **Traffic measurements:** they are related to SDCCH, TCH, etc.
- **Resource availability measurements:** availability of TCHs and SDCCHs, congestion time of TCHs and SDCCHs, average and peak number of busy TCHs and SDCCHs, etc.
- **Resource Access measurements:** number of messages sent in Abis interface, average load of control channels, etc.
- **Handover measurements:** number of successful/unsuccessful HOs, HOs per cause, etc.
- **Quality measurements:** number of radio measurements in the UL or DL path with quality in a certain range.

For practical purposes, several counters are usually combined to provide a meaningful performance measure. This is the reason why *Key Performance Indicators* (KPIs) are defined by network manufacturers in order to allow more efficient performance monitoring (see Section 4.4). KPIs are calculated using formulas for the counters in the NMS. Typically, KPIs describe the success/failure rates of the most important events such as handovers or dropped calls. An example of KPI is the percentage of handovers due to bad quality in the DL path, which can be calculated based on individual counters as:

$$DL\ Qual\ \ HO = \frac{sum(HO_{DL\_ql})}{sum(HO)} \tag{2.1}$$

where HO_{DL_ql} is the number of HOs in a measurement period (i.e. an hour) due to bad quality in the DL and HO es the total number of HOs in an hour. The *sum* was defined over all the hours of a day, so that KPIs are obtained per day.

Quality can be measured by means of different KPIs:

- **Mean Opinion Score (MOS):** MOS values range from 1 (bad) to 5 (excellent). In order to calculate MOS, a group of listeners should evaluate the perceived subjective voice quality under different conditions.

- **Bit Error Rate (BER):** BER measures the raw bit error rate in reception before the decoding process takes place. In GSM, BER is mapped into bands (i.e. RXQUAL, see Chapter 4).

- **Frame Erasure Rate (FER):** FER is highly correlated with the voice quality perceived by an user. FER represents the percentage of frames being dropped because the receiver is unable to correct bit errors in the most important part of the speech frame. The traditional problem associated with the usage of FER in GSM systems has been the lack of FER measurements in the NMS, although very recently operators have started to include them in their statistics.

- **Dropped Call Rate (DCR):** DCR measures the percentage of lost connections. This KPI is very important because a dropped call has a very negative impact on the quality of service perceived by the end user. There are different ways to calculate DCR: dropped calls per erlangs, dropped calls over originated calls, dropped calls over all calls handled by each cell, including incoming handovers, etc.

- **Call Setup Success Rate (CSR):** CSR measures the success rate of the signalling process related to setting up a call.

Apart from KPIs, alarms are generated at several points of the network to indicate a failure. Subsequently, they are transmitted and stored in the NMS. Although alarms are symptoms of malfunctioning, they do not necessarily point to the exact cause of the problems. This is due to two main reasons. Firstly, a single fault normally triggers more than 100 alarms that are often difficult to interpret. Secondly, a single alarm may be a manifestation of multiple faults or even be triggered in the absence of a fault.

2.2 State of the art

2.2.1 Automation and optimization

The coming years will bring profound changes to the mobile telecommunications industry [195]. Those changes are due to the new services offered by mobile operators, the introduction of new air interface technologies and the increased level of competition. It is foreseen that in the future, mobile phones will be the preferred choice for connecting to the Internet. This change is driven by the desire of the end users to have Internet access at any time regardless of their location.

To enable the convergence of mobile systems with the Internet, new functionalities will have to be integrated into existing cellular systems. For GSM operators, this has meant adding GPRS related hardware and software to existing networks. Similarly, the core / transmission network will have to be upgraded to facilitate data throughput for the new services. In parallel, operators are deploying UMTS networks in order to enable new services that require even higher data rates.

The addition of new technology increases the complexity of operating the network. When speech telephony was the only service offered, indicators such as DCR, speech quality and blocking of call attempts were suitable indicators to give operators an understanding about the performance of the network. With the addition of the packet-switched service, new performance indicators such as data transfer speed, delay or availability are required. Thus, the increased complexity of the network makes its analysis and tuning processes far more difficult.

The huge fees operators had to pay for UMTS licences have increased their debt burden significantly. In addition, further investment is needed to purchase the UMTS equipment and finance its installation. Consequently, operators are under increasing financial pressure to produce returns on their investments. At the same time new 3G operators have been admitted into several markets hence further increasing competition in the mobile industry.

These changes of the network go along with a continuous network roll out even in relatively mature cellular markets. The deployment of additional sites is mainly required to provide increased levels of in-building coverage as mobile users expect to be offered service in all geographical locations. In addition, network growth is also needed to enhance system capacity in order to cater for the rapid increase in the level of packet-switched traffic.

In the past, operators managed to cope with rapid technological changes and growth of the network by increasing their workforce. However, due to the financial pressures this is not a viable strategy anymore. Therefore, the only feasible option to maintain network quality with the existing workforce whilst also integrating new technology into the network is to increase the level of automation. This will free resources, which can tackle upcoming challenges brought about by the continuing evolution of the network.

Areas that require automation to offer a cost-effective implementation are those involving heavy calculations and/or the evaluation of several input parameters. In contrast to humans, computers can easily carry out analysis of complicated input data. In addition, computers are able to repeat a method many times and thus enable the application of the method to all cells in the network. The benefit of automation to these kinds of tasks is an increase in network performance since without the application of automation these tasks could not be carried out on a wide scale, i.e. a large number of cells. Automation can also be applied to existing tasks, e.g. frequency planning. In this case, one positive side effect of increasing the level of automation is that the staff is freed from mundane and tedious tasks and can apply their skills in new areas, which bring additional value to the operator. In this sense, automation assures that the staff is working on challenging tasks, which in turn helps to motivate and retain staff.

Operational efficiency is related to the work-effort that is required to provide a given spectral efficiency[10] at a given grade of service, which is crucial in providing cost-efficient data transfer. Several factors have influence on operational efficiency, such as the salary of the employees. For example, if a given feature enables an increase of spectral efficiency by 10%, but the related work-effort has to be increased by 50% to enable this improvement, then it is doubtful that the overall data transfer cost can be reduced by this feature.

[10]The *spectral efficiency* is a measure of how effectively a system makes use of a limited amount of spectrum. It can be defined as the delivered bits per second per Hz of occupied bandwidth

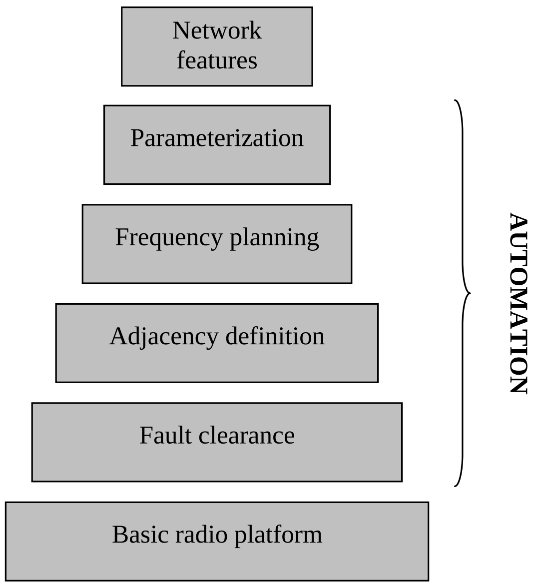


Figure 2.3: Factors contributing to the performance of the RAN of a cellular network

Automation helps to increase operational efficiency, while also increasing network performance. Thus, it is possible to obtain better performance from the existing network at a lower cost. Although the technical focus on this thesis is on RAN in GSM, the trend of increasing automation in cellular systems is also applicable to 3G networks.

Fig.2.3 displays several technical factors contributing to the performance of the RAN of a GSM network and their interdependence.

The foundation of a well-performing network is the **basic radio platform**. Issues such as the location of the sites as well as positioning of antennas [196] have a major impact on the distribution of the radio signal and, thus, on call quality. For example, the closer sites are located near a traffic hot spot because the easier it is to provide good call quality with low signal levels, hence reducing interference to surrounding cells.

In addition, the **clearance of faults**, called *troubleshooting* (TS), is also very important in order to ensure that the network operates exactly as designed [39, 43]. If a cell is temporarily non-operational, the performance of all cells in the vicinity is impaired. Ensuring that this cell is speedily brought back into operation is a crucial task to ensure optimum network performance.

The **adjacency plan**[11] is one of the major contributors to network quality. HO to most suitable cells will be possible only with the definition of correct adjacencies. On the contrary, without the definition of the optimum adjacencies, HOs occur to cells that are suboptimal in terms of their radio link performance and subsequently dropping of calls is very probable. Automating the process of updating adjacency definitions by deleting under-used ones and

[11]For handovers (HOs) to occur between a pair of cells, the target cell has to be defined as an *adjacency* at the source cell. If an adjacency is not defined, a HO cannot take place even if both cells are very close in propagation terms. The adjacency plan is the list of adjacencies defined for each cell in the network.

creating missing ones [186] will enhance operational efficiency and, depending on the quality of the existing adjacency management process, also increase network performance.

Similarly, a good **frequency plan**[12] is of paramount importance for network performance. The better the frequency plan, the less co- and adjacent-channel interference is experienced by subscribers. This means better call quality and subsequently smaller DCR. Thus, automation may help to improve the frequency plan and the work-effort associated with its creation, i.e. enhancing operational efficiency [38, 194].

Another important factor that determines network quality is **parameterizations of network features**. Traditionally, standard parameters have been used network wide, but this solution is not optimal when trying to enhance network performance. Each cell operates under unique conditions and requires customization of parameter values to provide optimum performance. In this context, automation is a very useful tool to increase operational efficiency [185, 195, 125].

Automation can be applied to the tasks indicated in Fig.2.3 in order to enhance network performance while reducing the manpower required to carry out the task. It is worth noting that information to tune the network is provided for free, since most of it is made of measurement reports that ordinary subscribers are providing free of charge in call mode. More information about automation and optimization can be found in [134, 135, 113, 177, 90, 125].

2.2.2 Troubleshooting in current cellular networks

As networks evolve and they increase in size, complexity and heterogeneity, the need for advanced fault management capabilities becomes critical. The RAN of cellular systems is no exception. Fault management, also called troubleshooting, includes the detection, isolation and correction of faults, where a fault is a cause of malfunctioning.

As described in Chapter 1, currently, in most cellular networks, TS is a manual process carried out by *experts* in the RAN. This TS process is characterised by eliminating likely problem causes in order to single out the actual one. During the procedure, several applications and databases have to be queried to analyze performance indicators, cell configurations and alarms. The speed of identifying faults is dependent on the level of expertise of the troubleshooter, the type of information available and the quality of the tools displaying relevant pieces of information. This means that, in addition to a good understanding of the possible causes of the problems, a very good understanding of the tools available to access the sources of information is also required. Due to the complexity of the management system, it is almost impossible for newcomers to perform TS in a proficient manner. Therefore, some of the most experienced members of staff are involved in rectifying problems in the network rather than looking after its further development. In addition, normally the experience on how to identify problems is kept "secret", every expert follows his own rule of thumb and if he leaves the company, his experience is lost. Furthermore, the growing size of cellular networks, together with their increasing complexity, make it very difficult for humans to analyze the vast amount of information coming

[12] The *frequency plan* is the list of frequencies assigned to each cell in the network

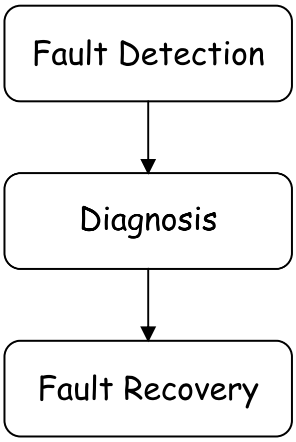


Figure 2.4: Phases in troubleshooting

from the network. Consequently, several operators consider TS as one of the most time and resource consuming tasks associated with the operation of the network. When downtime of cells is long, it unavoidably translates into lost of revenues for the operator. Thus, operators aim to limit the occurrence of faults as well as their duration.

Troubleshooting phases

TS in a cellular network consists of the following phases (Fig.2.4):

1. **Fault detection:** malfunctioning cells should be identified based on the values of KPIs and alarms (e.g. isolating cells with a high number of dropped calls).

2. **Diagnosis:** the cause of the problems (e.g. interference) should be identified based on KPIs, alarms and configuration data.

3. **Fault recovery:** some actions should be carried out in order to solve the problems (e.g. improving the frequency plan).

This thesis is focused on diagnosis, which is by far the most difficult and thus time-consuming task within the TS activity.

Operational Scenario

Currently, troubleshooting is mainly a manual process carried out by different teams. Each network operator follows its own approach, although Fig.2.5 is in line with TS in most networks.

TS is normally related to a *Trouble ticket* (TT) system. When a fault is investigated, a TT reflects the problem status, that is the fault description and the steps performed so far to solve it out or the identification of the faulty equipment in case the problem is deemed as a HW fault. A TT system is deployed as a large database, which can be queried by the user, using criteria like time constraints or identifiers (ID), such as cell ID and site ID. After a query, all the cases related to the specified cell within the given time period are shown. The entries are normally in "free text", almost like a "virtual log-book" that everyone uses to annotate the actions taken and observations made for the cell. Hence, in difficult cases lots of people from various departments have looked into the cell and potentially applied some changes. For example a case might involve changing parameters, swapping HW, re-tuning the cell, taking more measurements, re-tuning neighbouring cells, etc. Thus, several notes are normally written to such a case.

TS procedure may involve several layers, depending on the complexity of the case. The Front Office group (FOG) is the first such layer. This team is responsible for dealing with alarms generated by the network. Typically, the number of triggered alarms may be up to 100000 per week for the whole network. Alarm reduction technique reduces this large number of alarms so that the FOG are presented with the important alarms, leaving about 600 alarms per week. Then, FOG raises TTs, i.e. reports related to those alarms. In addition, FOG carries out a preliminary analysis to identify the cause of the problem. Sometimes this group solves the problem. In that case, they update the TT with a description of the performed actions and "close the TT". In more complex cases, they send the TT to the second layer for further investigation, including in the TT a description of the steps that were carried out.

The second TS layer is the Back Office group (BOG), which looks into faults based mainly on short-term statistics. BOG staff pursue a deeper analysis to identify the cause of the problem and they update the TT with the executed actions. If, as a consequence, the problem is solved, they close the TT. If they are not able to solve the problem, they reassign the TT to a more specialized group, the Technical Support group (TSG). This new group may be composed of troubleshooting experts of the operator itself or external experts, e.g. from a equipment manufacturer.

Therefore, TSG is the subsequent layer in the procedure. This team often uses scripts to generate a list of "worst performing cells" (i.e. they use fault detection tools) and takes this as a starting point for their work, focusing on the most serious cases. Then, they look further into TTs raised by FOG and BOG and raise new TTs with the results of their analysis. They can solve parameter related problems, but they often need to involve field engineers for problems related to HW on the site. Field engineers travel to the BTS sites and fix HW problems or any other problem requiring on-site personnel. Each day the field engineers receive a new plan containing the list of sites they have to visit and the cases they have to investigate on site. If the problem is solved at the site, TSG closes the TT.

Finally, a minor part of TTs are raised by the Customer Faults group (CFG) who may receive customer complaints from call centers, management, engineering staff, etc. They can either solve the problem themselves or send the TT to a more adequate group.

Trouble ticket systems are commonly used in telecommunication networks to assist in fault management [136, 139, 138]. Thus, TS begins with the documentation of a trouble in a TT. Subsequently, the trouble ticket system receives as inputs the reports generated by the different teams in Fig.2.5. Management of these reports implies the assignment of the tickets at each moment to the most adequate team to deal with it. The trouble ticket may pass through several hands and undergo different degrees of escalation with respect to priority until it is closed, i.e.

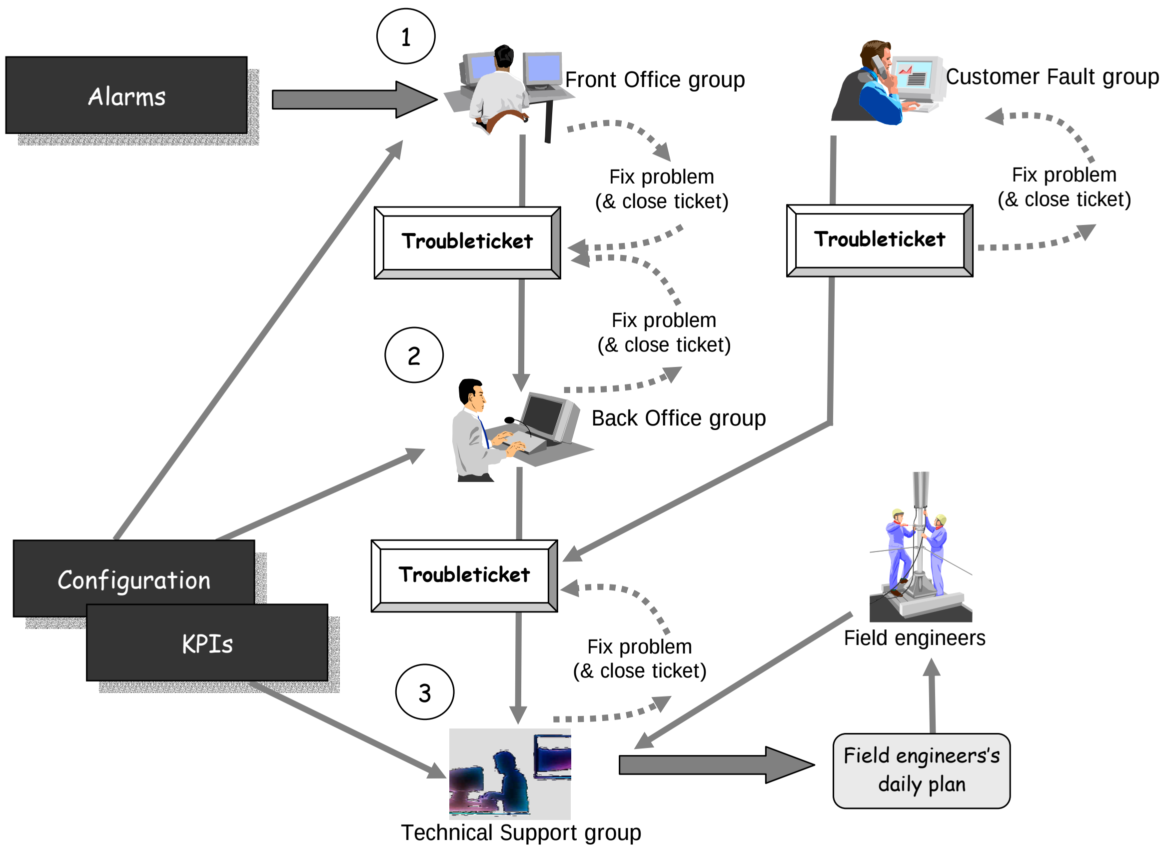


Figure 2.5: Troubleshooting in current mobile communication networks

the problem is solved.

Visualization tools

As explained above, in current cellular networks, diagnosis is carried out manually by experts with a lot of experience in the problem field. In such a case, the most informative portion of the measurement data should be represented in an efficient form in order to successfully exploit the expert knowledge. Hence, cellular network operators have visualization tools which are used for configuration, fault and performance management. The user can either analyze some predefined KPIs or he can specify new formulas to calculate KPIs from counters in the NMS. Those tools allow not only the observation of some instantaneous values of KPIs, but they also show the temporal evolution of KPIs, which is very useful in trend analysis. In addition, it is possible to define thresholds for KPIs in order to detect abnormal behaviors of network elements. Three types of reports are normally used: alarms, network configuration or measurements.

Experts in diagnosis use those visualization tools in order to identify faults. On the one hand, these experts can observe the main alarms. Alarms are graphically displayed on the screen in such a way that it is easy to identify the element that triggered the alarms in the first place. On the other hand, several KPIs related to the same fault can be represented together on the same chart. Experts analyze trends and abnormal behaviors of KPIs. Trends are useful to perform proactive fault management, that is identifying and solving a future fault before it becomes a real problem. Abnormal behaviors are sudden changes of KPIs, which indicate that a fault started at certain moment in time. In such case, reactive TS is crucial for performing fast fault identification and recovery so that network downtime is minimised.

Fig.2.6 shows some examples of the type of displayed information, which is analyzed by experts in order to diagnose a fault. In the figure some charts related to traffic channels, handovers due to bad quality and received signal quality in the downlink and uplink paths are depicted. From the graphics, it can be concluded that from April 30th there was a problem related to the uplink path.

Sometimes, operators have a kind of recipe describing the KPIs that should be observed in order to identify a certain fault. Instructions may come in the shape of a flowchart. For example, if the number of level handovers is high, then the following step proposed by the flowchart could be to check the received signal level. In other case, the signal quality should be analyzed. Nevertheless, more often troubleshooting experts follow no written guidelines, but their own diagnosis rules, which are based on their experience of working in this field.

2.2.3 Automatic troubleshooting

As described in Chapter 1, automating TS has several benefits. Firstly, the time required to locate the reason for a fault causing a problem is greatly reduced. Thus, the downtime and the time with reduced quality of service (QoS) is limited significantly. Secondly, fewer personnel and, thus, fewer operational costs are necessary to maintain a network of a given size at a given performance. Thirdly, the TS process is de-skilled as the majority of problems can be rectified

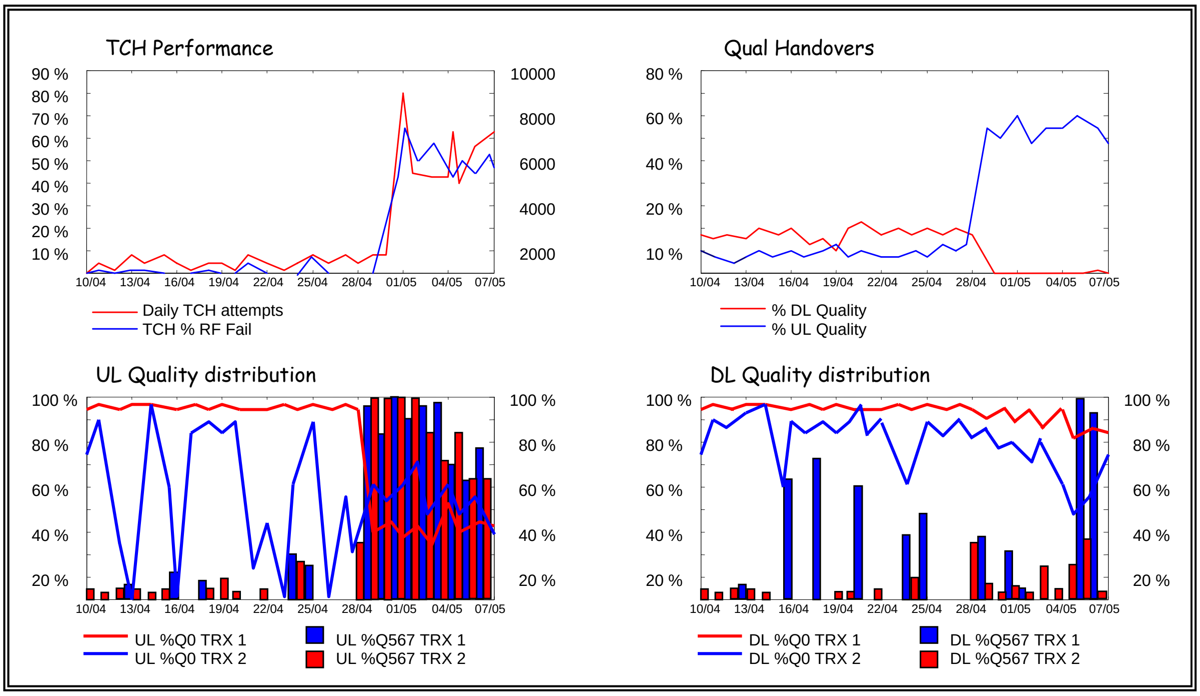


Figure 2.6: Examples of KPI visualization charts

with the help of the automated TS tool. Finally, one additional benefit is that the TS knowledge can be stored in the TS tool, therefore not being dependent on staff availability. Thereby, the gains achieved thanks to automated TS for an operator are significant as fewer personnel with a lower skill level can solve more network problems in less time. In the following section, a scheme to perform automatic TS will be proposed.

Decision-theoretic troubleshooting

Sometimes, fault identification is an iterative process: if after executing a specific action the problem is not solved, the analysis to identify the fault is repeated taking into account that the erroneous cause is discarded from the analysis. This is related to *decision-theoretic troubleshooting* [96, 51, 189, 111, 110], which considers that the aim of TS is to solve the problems, not just determining the cause.

At any stage of the TS process, there are many possible pieces of evidence or tests (*observations*), and repairs (*actions*) that can be collected or applied, respectively. Because these operations are expensive in terms of time and money, it is desirable to generate a sequence of *steps* (observations or actions) that, whilst minimising costs, results in a functioning network.

A very simple TS procedure collects all observations. Then, it makes reasoning based on the observations, ranking the possible causes according to their probabilities, and it performs the repair action associated with the most probable cause. If after executing this repair action, it is observed that the most probable cause was not the actual cause, the repair action associated to the following cause in the list of ranked causes is carried out. This is not an optimum procedure because all observations are collected, when probably not all of them are required to find out the cause. This is specially important in the medical domain, where some tests may be expensive and invasive for the patient. Furthermore, sometimes it could be more efficient to perform an action different to the one associated to the most probable cause. For example, suppose that the most probable cause was a hardware fault. In that case, it is probably more cost-efficient to make a simple parameter change to check whether the problem was a configuration problem than replacing the faulty HW unit. Only if the parameter change fails to solve the problem, someone will be send to the site to change the problematic HW unit, which is always a very expensive and time-consuming action.

An optimum procedure would order the steps (observations and actions) in a sequence minimising the cost (money and time). The actions would be ranked according to their efficiency, where the *efficiency* of an action is defined as the ratio between the probability of the action solving the problem and the cost of that action. At each moment, a greedy algorithm would calculate the expected cost of the sequence of steps carried out up to the moment together with each new candidate observation and action. The chosen step would be the one with the lowest expected cost. Following this algorithm, in general, some actions will be carried out before collecting all observations and very probably the problem will be solved by one of these actions, saving time and money. For example, if after collecting some observations, the probabilities of faulty HW and bad configuration are very high, a parameter change will be proposed by the algorithm and if the problem is solved, it will not be required to collect the other observations or perform the other actions.

In the problem under study in this thesis, it has been considered that all observations (performance indicators, alarms and configuration data) are available in the NMS and the cost of collecting them is negligible. In future work, the time saving of using only some of the observations could be quantified. In addition, out of the three TS tasks in Fig.2.4, diagnosis is the most difficult and time consuming and no information about it can be found in present literature. Hence, this thesis is focused on diagnosis. It is supposed that there is a Fault Detection module which identifies the problematic cells prior to diagnosis. Under these assumptions[13], the optimal repair sequence for a faulty cell is given by the following algorithm:

1. Collect all available observations for the cell under study
2. Given that the cell is malfunctioning, compute the probabilities of the causes
3. Execute the (as yet non executed) action A_i with the highest efficiency $\epsilon(A_i|E)$:

$$\epsilon(A_i|E) = \frac{P(A_i = yes|E)}{C_{A_i}(E)} \tag{2.2}$$

   where $P(A_i = yes|E)$ is the probability of the action A_i solving the problem given the evidence compiled so far and the result of previous actions, E, and $C_{A_i}(E)$ is the cost of

[13] See [96] for detailed description of assumptions and demonstration.

performing the action A_i, which may depends on E. It is assumed that linked to each cause C_i there is a unique action A_i which solves the problem. Hence, $P(A_i = yes|E)$ is also the probability of the cause C_i.

4. If the action solved the problem, then terminate. Otherwise, go to step 2.

In this algorithm, the required information are the posterior probabilities of the causes given evidence and the costs of the related actions. The costs can be easily provided by operators of cellular networks. Therefore, cost will not further explicitly appear in this document, instead it will focus on calculating the probabilities of the causes.

Interfaces of the automatic troubleshooting system

In this thesis, the expert system which performs automatic diagnosis and proposition of healing actions will be denoted *Diagnosis and Recovery* tool (DAR). However, it should be understood that troubleshooting has a wider meaning because it is composed of fault detection, diagnosis and fault recovery, as described in Fig.2.4. The tool that pursues all those tasks will be denoted *Troubleshooting tool* (TST).

Fig.2.7 shows the interfaces of the DAR. The Fault Detection subsystem provides the DAR with the list of faulty cells to be diagnosed. DAR requires a *diagnosis model* on which its reasoning mechanisms are based. The subsystem named *model definition* is in charge of building the diagnosis model to be used by the DAR. Diagnosis models can be built based either on the expertise of human troubleshooters or on statistics from the network. Those statistics are normally saved in the NMS. The inputs to the DAR are configuration parameters, alarms and KPIs for each of the faulty cells. The NMS contains historical databases containing the values of all those inputs. In addition, DAR may also require some inputs directly requested to the user, which may be related to observations that are not in the NMS, e.g. whether the day was rainy on the day the fault occurred. The output of the DAR is a diagnosis on the fault that is causing the problems in each malfunctioning cell. In addition, DAR proposes a list of actions, ranked by their efficiency, to be sequentially executed until the problem is solved. These actions may be just changing a configuration parameter from a remote terminal or may involve sending personnel to a site to replace a faulty piece of equipment. A TST may even execute software related repair actions. Although, normally, operators prefer that the TST only proposes the actions, but the final decision be handled by a human expert (the TST in this case acts as a so called *decision support system*). Finally, DAR is also intended to generate a report about the diagnosed cause and the steps carried out in order to recover from the fault (trouble ticket).

The TST can work independently from the NMS, but most of the benefits of automated TS are achieved when it is an integrated part of it. This integrated solution will provide direct access to information required in fault analysis as well as access to the operators's fault management system. An integrated solution is also beneficial in case of TS of multi-vendor networks and of multi-system networks (GSM, UMTS, WLAN). The TS expert system is system independent, thus, with minor modifications multi-vendor and multi-system networks can be supported. The

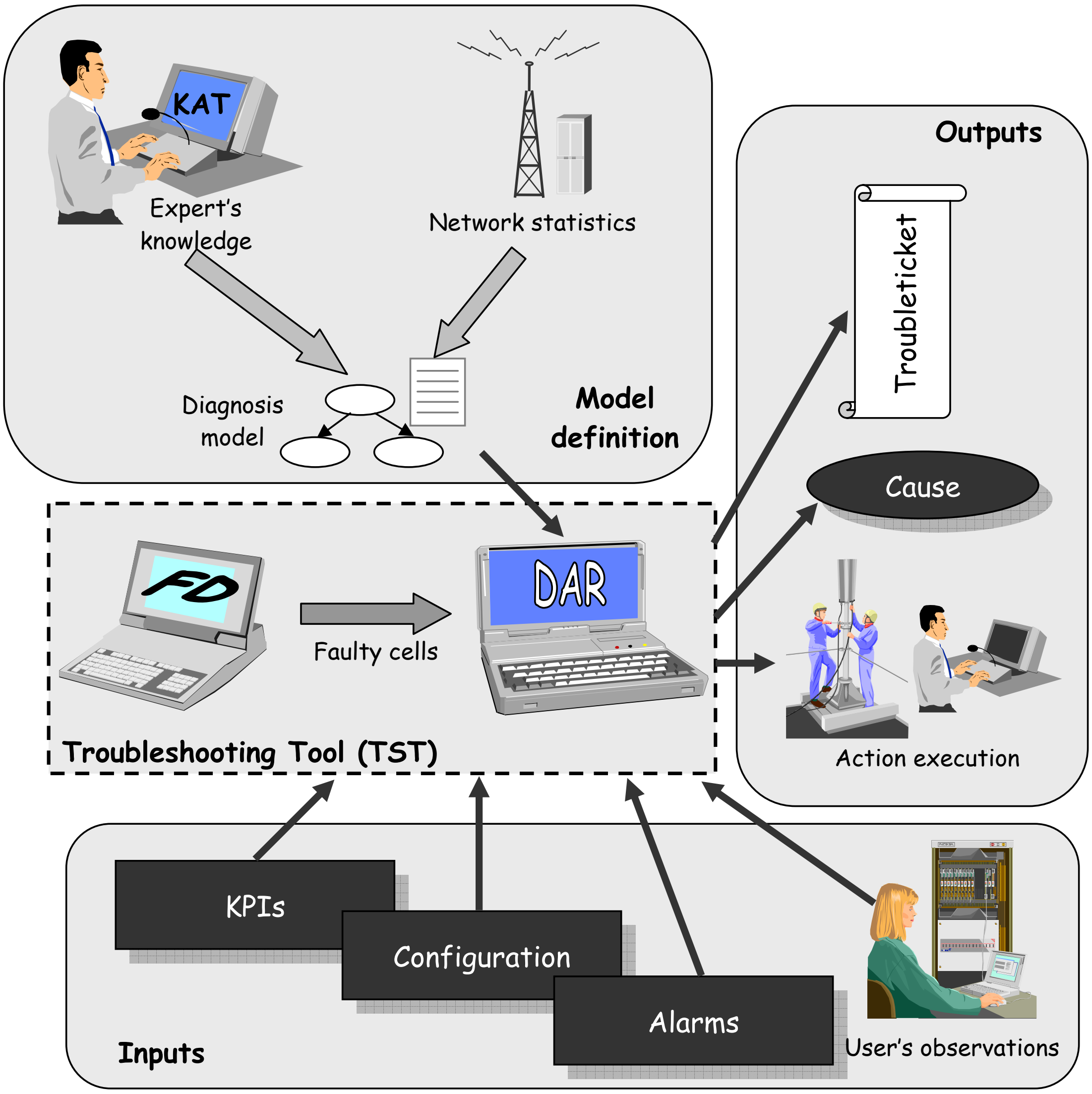


Figure 2.7: Troubleshooting system interfaces

integration in the NMS guarantees that the TST is whole synchronised with the whole fault management system. Hence, all relevant TS cases can be automatically directed to the TST and if it finds the solution, the case is cleared, reported and filed. If the problem is not found by the expert system, it can be redirected to the specialists for further analysis and the final conclusions can be incorporated into the knowledge of the expert system.

First steps in automation of troubleshooting

Preliminary studies related to automatic troubleshooting have been focused on methods to achieve efficient visualization of the network performance. Thanks to those visualization tools, anomaly detection is carried out more easily. For example, in [132] a method to analyze network performance based on knowledge and standard quadratic programming techniques is presented. The basic idea of the method is to measure performance in terms of the number of failed operations in the network. Thus, KPIs are divided into sets describing the performance of the subsystems of the GSM network. Simple mathematical models are proposed for each subsystem, which depend on the KPIs. The model parameters are estimated from training data using quadratic programming. After the model is estimated, a graph may be constructed in order to analyze the dependencies between the subsystems and the KPIs.

Several studies have been carried in the field of Fault Detection, which is the first step in TS, as shown in Fig.2.4. Most methods consist in building models for the normal behavior of the system. The deviations of the available measurement variables from the normal behavior can then be detected with some type of abnormality detector.

Hence, in [123, 124, 104, 133] the proposed method for FD is based on a Self-Organizing Map (SOM), which is one of the most popular neural network[14] algorithms due to its efficient visualization properties. A behavior pattern of a cell is a set of KPIs. SOM maps a high-dimensional behavior patter onto a low-dimensional, usually two-dimensional grid. In network analysis, SOM can be used to find and show similarities between behavior patterns of cells. Thus, when SOM is visualized, similarly behaving cells can be spotted close to each other. In these papers, clustering, that is grouping of similar behaving cells, uses a combination of SOM and an algorithm called k-means. Cell clusters that have been found by the proposed methods can be used to identify cells that have a certain defect, e.g. by calculating the distance between a new input vector to the normal profile.

In [48, 47], a method for FD in 3G cellular networks is proposed. A neural network is trained using an algorithm called *Winner-Take-All*, in which only the weight of a single neuron is changed with each new input vector. The neural network is trained with vectors of KPIs collected during normal functioning of the cellular network, i.e. no examples of abnormal features are used for training. Using percentiles of the normality profile, a numerical interval representing normal behavior of the system is defined. FD is carried out in the following way: the defined interval for normal behavior is used to classify a new vector into normal/abnormal by means of hypothesis

[14] A neural network is a type of artificial intelligence method that attempts to imitate the way a human brain works. A neural network comprises several interconnected elements, which process information simultaneously, adapting and learning from past patterns.

testing. Furthermore, in these papers, the method is extended to identify anomalous attributes. Thus, confidence intervals for the normal behavior of each KPI are calculated. Each new KPI value is compared with the normal behavior and, if it is not within the range defined for the normal interval, then it is considered abnormal. The assertion of a KPI being abnormal may be taken as input for a more complex diagnosis system where symptoms have been discretized into two states: normal / abnormal.

Regarding automatic diagnosis, up to our knowledge, no references can be found about diagnosis in the RAN of cellular networks. However, automatic diagnosis has been extensively studied in other fields, such as diagnosis of diseases in medicine [34, 146, 99, 199, 35, 154, 147, 144], troubleshooting of printer failures [96, 51, 100, 94, 98, 101, 174, 111, 175, 173], diagnosis of faults in satellite communication systems [107, 131, 44, 57], etc. Fault identification in the core of communication networks has also received considerable attention [115, 182, 190, 73, 183]. In that scenario, it is very important to represent the dependencies among communication system entities because a failure in an entity may have a large impact on other system entities. Furthermore, in general, the only type of symptoms are the alarms related to the network entities. In the RAN of cellular systems, the previous formulation is not valid, because of two main reasons. Firstly, most faults are not related to a physical component, but to a poor network planning or an incorrect parameter setting. Thus, modelling the interactions among entities is not a requirement like in other communication networks. Secondly, although alarms are very important to identify faults in specific pieces of equipment, they do not provide conclusive information in order to isolate more general problems, such as radio frequency interference or lack of coverage. In this scenario, performance indicators are crucial to identify the fault. Furthermore, the fact that performance indicators are continuous, instead of binary like alarms, introduces a new difficulty in the modelling which is nonexistent in the case of other communication networks.

As described in Section 2.2.2, in current cellular networks, diagnosis in the RAN is still a manual process, although very recently research studies in automation have been initiated. References in this area are focused on alarm correlation. In [84, 88], model-based alarm correlation methods are presented. These approaches use a model of each device formulated as a set of formulas. For example, a certain device is considered to generate a malfunctioning alarm if a specific set of low-level alarms are present. In [193, 192], a neural network based alarm correlation system is proposed. On the one hand, each generated alarm is represented as a neuron at the input layer. On the other hand, each filtered alarm is represented as a neuron in the output layer. During training, the weights in the neural network are adapted.

Although alarm correlation can be considered a first step in the diagnosis of faults, alarms do not provide conclusive information to identify the cause of problems, especially if the possible causes are not only faults in pieces of equipment. Even after alarm correlation, the number of triggered alarms for a single cause is normally very high. In addition, the same alarm can be triggered by different causes. Because of these reasons, it is impossible to identify all fault causes based only on alarms. Faults, such as interference or lack of coverage are difficult to identify if performance indicators are also not considered.

This thesis is focused on automatic diagnosis based on performance indicators because it has not been previously addressed in the existing literature and because it is the main part of the current TS process applied by operators around the world. Nevertheless, an efficient diagnosis system should consider not only performance indicators, but also alarms.

Chapter 3

Diagnosis techniques

The first part of this chapter summarizes some techniques which may be used to model uncertainty in reasoning. The second part of the chapter is devoted to the principles of BNs, which in this thesis, have been the selected method amongst the previously described ones.

3.1 Reasoning under uncertainty

3.1.1 Introduction

Uncertainty is always intrinsically joined to diagnosis. When an expert asserts which was the cause of problems in a cellular network, he/she is never completely sure about his diagnosis. There are different sources of uncertainty. Firstly, the data could be unreliable. For example, a measuring equipment may be defective. Secondly, the data may be incomplete. For example, information about only some of the symptoms may be available. Finally, the data may be only approximately known. For example, a signal level may be measured with a limited degree of precision. Furthermore, not only might the data be imprecise, but so might be the rules for drawing conclusions. That is, the knowledge is not deterministic. For example, the same symptoms may be related to different causes. Therefore, diagnosis requires a means for reasoning with uncertainty.

Different approaches to model uncertainty can be found in artificial intelligence bibliography, which will be briefly summarized in the following sections [181, 166]. Most techniques use the theory of probability to deal with uncertainty. In the literature, a philosophical debate about the meaning and interpretation of probabilities can be found. There is a distinction between *objective probability*, which is linked to the convergence of a relative frequency of an experiment, and *subjective probability*, also called *degree of belief*, which is an individual's subjective estimate of the certainty of an event. For example, the experiment of tossing a coin could be repeated many times and the "objective" probability of any of the outcomes would be the limit, as the number of trials approach infinity, of the relative frequency of that outcome. However, if a person has to bet that A will be the winning team on a given upcoming football game, the probability would be "subjective" because the game cannot be repeated many times under the exact same conditions.

It should be pointed out, that a fact is either true or false (i.e. either team A will win or not). Thus, the degree of belief is different to the *degree of vagueness*, which is used to define concepts that imply vagueness. For example, Maria is 0.7 tall would indicate the degree (from 0 to 1) in which Maria is tall.

3.1.2 Certainty factors

Certainty factors (CFs) were used to develop MYCIN [172]. The MYCIN system, which aimed at diagnosing and recommending treatment for certain blood infections, was created at Stanford in the 1970s. In this system, the model was composed of facts and rules[1]. CFs linked degrees of belief to those facts and rules. On the one hand, the CF associated with a hypothesis[2] stands for the degree of belief in that hypothesis, given all the evidence that has been used so far. On the other hand, the CF linked to a rule indicates the degree of belief in the conclusion following from the rule's premise. The CF is explained in terms of a subjective belief in a hypothesis and, thus, it ranges from 0 to 1. An example of rule is the following:

IF the received signal level is low

and the received signal quality is bad

and the average timing advance is high

THEN the cause of the high dropped call rate is a lack of coverage (0.8)

This rule states that if the conditions about signal level, signal quality and timing advance are met, then we can be 0.8 certain about the cause of the problem, i.e. the probability of lack of coverage is 0.8.

Some schemes have been developed to combine the CFs of different hypothesis and rules [52]. However, in [92], it was demonstrated that, when combining different rules, CFs could provide incorrect degrees of belief, due to the unavoidable inconsistencies of any model with many rules.

3.1.3 Dempster-Shafer theory

The Dempster-Shafer (D-S) theory [72, 169] is a mathematical theory of evidence[3], which distinguishes between lack of belief and disbelief. For example, let A represent the proposition "Maria is tall". Then, according to the axioms of probabilities[4] $P(A) + P(-A) = 1$. But if we do not even know who Maria is, we cannot say that we believe or not believe the proposition. It would therefore be meaningful to denote our belief of A, $B(A)$, and of $-A$, $B(-A)$, as both being 0. The main advantage of D-S approach over other techniques is its ability to admit partially specified models.

[1] Rules are IF-THEN assertions

[2] hypothesis are facts that are not observable or only observable at an unacceptable cost

[3] Evidence is the assignment of values to variables

[4] $-A$ is the complementary proposition of A, e.g. Maria is not tall

First, a complete set of elements should be defined, which is called the *frame of discernment* and it is denoted by θ. The elements in θ are mutually exclusive. For example, in diagnosis applications, θ would be the set consisting of all the possible faults or diseases. The set of subsets of θ is denoted by 2^θ. For example, if θ is the set {A (interference), B (lack of coverage), C (hardware fault)}, then 2^θ is the set $\{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\}$. The D-S approach uses a number between 0 and 1 to indicate a degree of belief. In this way, there is an assignment function, M, called *basic probability assignment*, which assigns a number $[0, 1]$ to every subset in 2^θ. Furthermore, M is such that the sum of assignments over all subsets is 1, i.e. $\sum_{\forall x \in 2^\theta} M(x) = 1$. $M(\theta)$ is a measure of the total belief that is not assigned to any subset of θ. For example, let $\theta = \{A$ (interference), B (lack of coverage), C (hardware fault), D (transmission fault)$\}$ be possible causes of the high dropped call rate in a cell. Suppose that we know that the cause is not lack of coverage to the degree 0.4. Then, we would assign $M(\{\text{A,C,D}\}) = 0.4$, $M(\theta) = 0.6$ and $M(x) = 0$ to the rest of subsets in 2^θ.

From the basic probability assignment, the D-S approach defines three other measures:

- The *belief function* $\text{Bel}(z)$, which assigns to every subset z of 2^θ the total belief assigned by M to z plus the belief assigned to all of its subsets, that is, $\text{Bel}(z) = \sum_{x \subseteq z} M(x)$
- The *measure of doubt*, $\text{D}(z) = \text{Bel}(-z)$, where $-z$ is the complementary set of z.
- The *measure of plausibility*, $\text{Pl}(z) = 1 - \text{D}(z)$. This measure is also called the *upper belief function* or the *upper probability function*.

$\text{Bel}(H)$ gives the total amount of belief committed to hypothesis H based on the available evidence, and $\text{D}(H)$ gives the amount of belief committed to its negation. $\text{Pl}(H)$ expresses how much we should believe in H if all currently unknown facts were to support H. D-S theory also defines how to combine degrees of belief using Dempster's rule.

Example. Let's assume $\theta = \{A, B, C\}$, $M(A) = 0.3$, $M(A, B) = 0.2$ and $M(A, B, C) = 0.5$. Then, $\text{Bel}(\{A\})$=0.3, $\text{Bel}(\{A, B\})$=0.5, $\text{Bel}(\{A, C\})$=0.3, $\text{Bel}(\{A, B, C\})$=1 and $\text{Bel}(\{B\})$=0. Consequently, $\text{Pl}(\{A\})$=1, $\text{Pl}(\{A, B\})$=1, $\text{Pl}(\{A, C\})$=1, $\text{Pl}(\{A, B, C\})$=1 and $\text{Pl}(\{B\})$=0.7.

3.1.4 Fuzzy logic

In the previously described approaches, certainty was referring to the degree of belief in something. For example, the sentence "I believe (0.8) it will rain tomorrow" does not mean that we expect 80% of rain. Either it will rain or it will not. Uncertainty refers to the probability of rain, that is the degree with which we believe it will rain versus the degree with which we belief it will not rain.

Fuzziness, which is a concept different to uncertainty, is related with vagueness. For example, let's consider a day in which it rains a few times during the day. Then, we could say that the day is rainy to some degree. Hence, fuzziness allows to define concepts that imply vagueness. It has nothing to do with degree of belief in something and does not have to be related to probabilities.

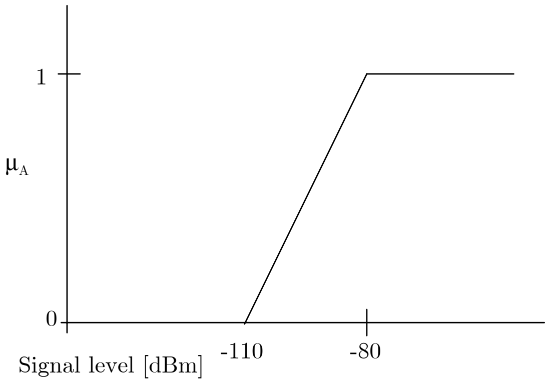


Figure 3.1: Characteristic function for a fuzzy set representing “high received signal level”

Formal theory about fuzzyness was first published in 1965 [205]. Since then, fuzzy set theory has been a burning field in mathematics and artificial intelligence. Fuzzy set theory is based on the idea of gradual membership in a set.

A *fuzzy set* A is characterized by a *membership function*, also called *characteristic function*, μ_A, which assigns each point x a real number in the interval [0,1], representing the degree or grade of membership of x in A. For example, Fig.3.1 shows a membership function which determines whether the average received signal level would be considered high. According to that function, any value under -110 dBm is not high ($\mu_A = 0$), whereas a value over -80 dBm is definitively high ($\mu_A = 1$). Received signal level whose value falls between these extremes are considered high to an intermediate degree. A variable like “level” is called a *fuzzy variable*, and it can take on values like “high”, “medium” or “low”. That is, a fuzzy variable takes on a fuzzy set as a value.

Fuzzy sets and characteristic functions may be used in two ways. Firstly, they can be used to estimate degrees of membership. For example, suppose we measure that the received signal level is -95 dBm and we want to determine the degree to which this measurement is a “high value”. A specific value like “-95 dBm” is called a *crisp value*. According to the characteristic function in Fig.3.1, we would say that the degree of membership is 0.5. Secondly, fuzzy sets can be used to express possibilities in situations where we have incomplete information. Suppose we are told that the received signal level is high but we do not know its exact value. In that case, the characteristic function is Fig.3.1 can be used to express preferences on possible values of a variable whose exact value is not known. This interpretation of fuzzy sets is called *possibility distribution*.

Operations on fuzzy sets are defined in a similar manner as operations on ordinary sets. Operators such as complement, union and intersection are applied to the characteristic functions [65]. For example, the complement of a fuzzy set is the set whose characteristic function is 1 minus the original set's characteristic function at each point. The union operation takes the maximum of the characteristic functions of two sets, whereas the intersection takes the minimum.

A *fuzzy relation* is a fuzzy set defined over a cross-product[5]. A fuzzy relation is represented by a characteristic function that associates a grade of membership with each element of the cross-product.

A *fuzzy proposition* is a statement that asserts a value for a fuzzy variable, e.g. "the received signal level is low". In this case, "the received signal level" is the fuzzy variable and "low" is the value for that variable, which is a fuzzy set.

A *fuzzy rule* relates two or more fuzzy propositions. Fuzzy inference must determine a belief in a rule's conclusion given evidence on the rule's premise, i.e. fuzzy inference specifies how to combine the fuzzy values with the rule so as to define a new fuzzy value. There are several fuzzy inference techniques, the most common ones being the *max-min* and the *max-product* inference. A fuzzy rule could be the following:

IF received signal level is low

and quality is poor

THEN interference is high

After the rules have been applied, the output is a fuzzy set, called *induced fuzzy set*. Sometimes, for example in control systems, this fuzzy set has to be converted into a crisp value. This process is called *defuzzification*. The most common methods for defuzzification are the *moments method* and the *centroid method*.

In summary, a reasoning system based on fuzzy logic would perform the following steps. Firstly, if the inputs to the system are crisp they must be converted into fuzzy vectors. Secondly, fuzzy rules are applied. Evidence is combined using fuzzy set union or intersection. Finally, the fuzzy set outputs of the system become crisp through defuzzification.

3.1.5 Probabilistic networks

A probabilistic network, also called *Bayesian Network* [157, 110, 54] is a pair (D, P) that allows efficient representation of a joint probability distribution over a set of random variables $U = \{X_1, ..., X_n\}$. D is a *Directed Acyclic Graph*[6] (DAG), whose nodes correspond to the random variables $X_1, ..., X_n$ and whose edges represent direct dependencies between the variables. An example of the DAG corresponding to a BN is depicted in Fig.3.2. The second component, P, is a set of conditional probability functions, one for each variable:

$$P = \{p(X_1|\pi_1), \ldots, p(X_n|\pi_n)\} \tag{3.1}$$

where π_i is the parent set of X_i in U[7].

[5] A cross-product of two sets, $X \times Y$, is the set containing all pairs (p, q) where $p \in X$ and $q \in Y$. For example, let $X = \{a, b\}$ and let $Y = \{0, 1\}$. Then $X \times Y = \{(a, 0), (a, 1), (b, 0), (b, 1)\}$.

[6] A DAG is a set of nodes and oriented edges with no paths that returns to the same node (i.e. no cycles)

[7] X_j is a *child* of X_i and X_i is a *parent* of X_j if there is a link from X_i to X_j. X_j is a *descendent* of X_i if there exists a directed path from X_i to X_j

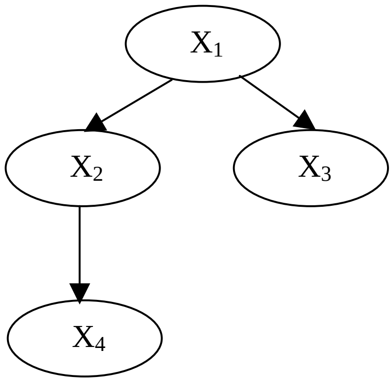


Figure 3.2: An example of BN

The set P defines a unique joint probability distribution over U given by

$$P(U) = \prod_{i=1}^{n} p(X_i | \pi_i) \tag{3.2}$$

BNs encode the conditional independence among variables. The edges of the graph represent the assertion that, given its parents, a variable is conditionally independent of its non-descendants in the graph. For example, in Fig.3.2, given X_1, X_2 is conditionally independent of X_3.

Variables may be continuous or discrete. If the variables are continuous the quantitative part of the BN is composed of conditional probability density functions (pdfs). On the contrary, if the variables are discrete, the quantitative part of the BN is composed of probability tables. *Evidence* $E = \{X_1 = x_1, ..., X_m = x_m\}$ is an assignment of values to variables in a subset of U, $X = \{X_1, ..., X_m\}$.

Belief networks may be used to obtain the probability of certain variable X_i given the available evidence, i.e. $P(X_i = x_i|E)$. This process is called *inference*, *evidence propagation* or *probability updating*. It is known that, in general, this task is NP-hard[8] [62, 66], although efficient heuristic algorithms have been developed [119, 69].

BNs will be further explained in Section 3.2.

3.1.6 Justification of the selected technique

Amongst the techniques described in the previous sections, Bayesian Networks have been the selected one. Up to our knowledge, there are no previous experiments on automatic diagnosis of RANs of cellular networks. Hence, our criterion is based on experience in other application domains and in general characteristics of the summarized techniques.

The main reasons for discarding the other alternatives are the following. Firstly, CFs are

[8] "Inference is NP-hard" means that it is not possible to obtain an algorithm of polynomial complexity for evidence propagation

not recommended, mainly because of the lack of solid theoretical bases, which can lead to incoherent conclusions. Secondly, as advantage, D-S theory accepts an incomplete probabilistic model when some parameters (e.g. prior or conditional probabilities) are missing. However, D-S theory is more difficult to interpret that BNs: it estimates how close the evidence is to force the truth of the hypothesis, instead of estimating how close the hypothesis is to being true. Finally, the philosophy of fuzzy logic is different to that of BNs, in the sense that fuzzy logic is concerned with vagueness, whereas BNs deals with probabilities. There are long discussions about the utilization of fuzzy logic vs. BNs [121, 79, 129, 197, 78]. Some authors argue that both techniques are complementary rather than competitive [204], whereas others defend that one of them includes the other [122, 128, 137]. In this thesis, probability has been preferred to fuzzy logic as mathematical basis to model diagnosis systems. However, in the future it is not ruled out to build a diagnosis system for cellular networks based on fuzzy logic and to compare results with the ones achieved in this thesis.

It should be pointed out that no diagnosis system other than the one based on BNs has been implemented. Hence, the selection of the technique has not been based on comparisons of the techniques above.

Although, up to our knowledge, there are no previous systems for diagnosis in cellular networks, the number of applications in other domains are enormous. The fields that have had more influence on this thesis are the following ones:

- Traditionally, medicine has been the outstanding application for automatic diagnosis systems. Hence, it is logical that BNs have been frequently used to build medical diagnosis systems. One of the first applications of BNs were Munin [34], which diagnosed neuromuscular disorders and Pathfinder [99], which performed diagnosis of lymph-node diseases. Since then, the number of applications of BNs to medicine have grown exponentially [199, 146, 35, 154, 147, 144].

- Due to their proximity, the application of BNs to telecommunications has also been the focus of our preliminary survey. BNs have been successfully used for fault management in the core of communications networks [115, 182, 190, 73, 183] or in satellite communication systems [107, 131].

- Diagnosis of faults in printers have also contributed to the development of troubleshooting based on BNs. The research carried out by Microsoft [96, 51, 100, 94, 98, 101] and HP [174, 111, 175, 173] are specially remarkable.

3.2 Bayesian Networks

In the following sections, the basis of BNs required to follow this thesis will be summarized. The approach in this introduction has followed [110]. For further study on BNs, readers are addressed to [110, 157, 54, 64].

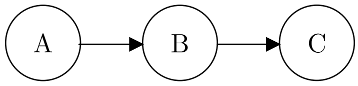


Figure 3.3: Serial connection

3.2.1 Introduction to BNs

A BN efficiently represents a joint probability distribution over a set of random variables. BNs are graphically depicted as directed graphs. A *directed graph* consists of a set of *variables* and a set of *directed links*, also called *directed edges*, between variables. In a directed graph, if there is a link between variable A and variable B, it is said that B is a *child* of A and A is a *parent* of B.

A variable may be continuous or discrete. In the latter case, the variable can have any finite number of states[9]. For example, a discrete variable may be the result of tossing a die (whose states are 1, 2, 3, 4, 5, 6) and a continuous variable may be the temperature in a room. A variable is in exactly one of its states, although which one it is may be unknown to us.

BNs may be used to infer how a change of certainty in one variable may change the certainty for other variables. For example, knowing that I have fever increases my belief in me having flu. When the state of a variable is known, it is said that it is *instantiated*. Relations among variables can be classified in the following types:

- **Serial connection.** In Fig.3.3, A has an influence on B, which has an influence on C. Hence, evidence on A will influence the certainty of B, which then influences the certainty of C. Similarly, evidence on C will influence the certainty on A through B. However, if the state of B is known, then the channel is blocked, and A and C become independent. It is said that A and C are *d-separated given* B.

  Example. A represents the state of the battery of a car (*OK/dead*), B specifies whether the car will start (*yes/no*) and C defines whether the car will move (*yes/no*). Knowing that the battery is dead ($A =$*dead*) strongly increases our belief in the car not moving ($C =$*no*). But if we know that the car does not start ($B =$*no*), knowing the state of the battery (A) does not give us more information about whether the car will move (C). BNs can be also used in the backward direction. In this case, knowing that the car does not move, highly increases our believe in the car not starting. In turn, if we believe that it is probable that the car will not start, we will increase our belief in the battery being dead. But if we know that the car does not start, knowing that the car does not move will not change our belief in the state of the battery.

  Rule 1: Evidence may be transmitted trough a serial connection unless the state of the variable in the connection is known.

[9]in this thesis, the terms "value of a variable" and "state of a variable" will be used indistinctly

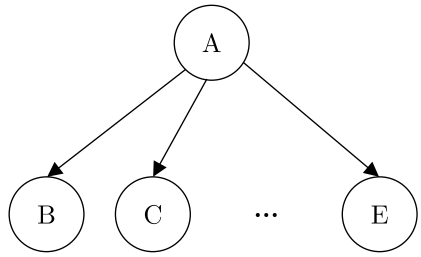


Figure 3.4: Diverging connection

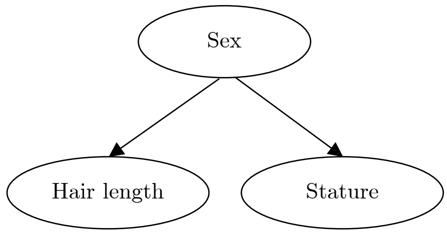


Figure 3.5: Example of diverging connection

- **Diverging connection.** In Fig.3.4, influence can pass between all the children of A unless the state of A is known. It is said that $B, C, ..., E$ *are d-separated given* A.

  Example. Fig.3.5 shows the causal relations between *Sex* (*male/female*), *length of hair* (*long/short*) and *stature* (< 168 *cm*/> 168 *cm*). If we do not know the sex of a person, seeing the length of his/her hair will tell us more about the sex, and this in turn will focus our belief on his/her stature. On the other hand, if we know that the person is a man, then the length of his hair gives us no extra clue on his stature.

  Rule 2: Evidence may be transmitted through a diverging connection unless it is instantiated.

- **Converging connection.** Fig.3.6 shows a converging connection. If nothing is known about A except what may be inferred from knowledge of its parents $B, ..., E$, then the parents are independent, i.e. evidence on one of them has no influence on the certainty of the others. In other words, knowledge of one possible cause of an event does not tell us anything about other possible causes. However, if something is known about the consequences, then information on one possible cause may tell us something about the other causes.

  Example. In Fig.3.7, if we do not know if a person has nausea or pallor, then the information on whether the person has a salmonella infection will not tell us anything about the person having flu. However, if we have noticed that the person is pale, then the information that he/she does not have a salmonella infection will increase our belief

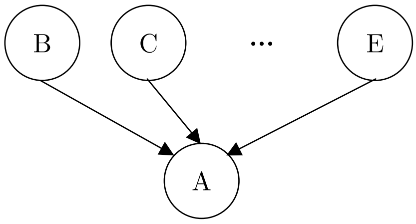


Figure 3.6: Converging connection

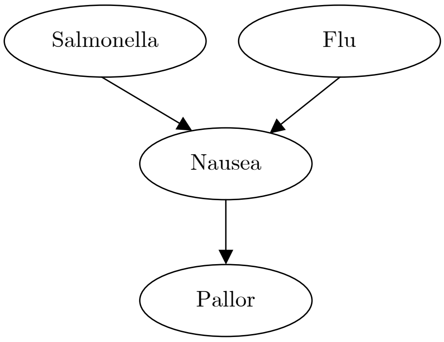


Figure 3.7: Example of converging connection

in he/she having flu.

Rule 3: Evidence may only be transmitted through a converging connection if either the variable in the connection or one of its descendants has received evidence.

Two distinct variables A and B in a causal network are *d-separated* if, for all paths between A and B, there is an intermediate variable V such that either:

- The connection is serial or diverging and V is instantiated, or
- the connection is converging, and neither V nor any of V's descendants have received evidence

If A and B are not d-separated, we call them *d-connected*.

Definition of BN

A discrete *Bayesian Network* consists of the following:

- A set of *variables* and a set of *directed edges* between variables
- Each variable has a finite set of mutually exclusive states

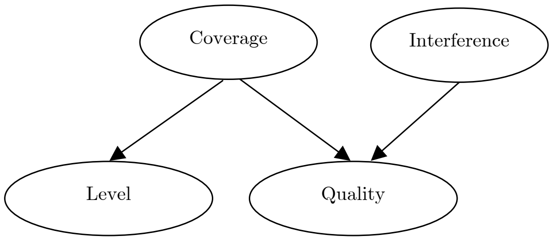


Figure 3.8: Example of BN

- The variables together with the directed edges form a *directed acyclic graph.*
- To each variable A with parents $B_1, ..., B_n$, there is a table[10] $P(A|B_1, ..., B_n)$ attached. If A has no parents, the table reduces to prior probabilities $P(A)$.

The definition of BN does not refer to causality, and there is no requirement that the links represent causal impacts. Instead, it is required that the d-separation properties implied by the structure hold.

Although discrete BNs are most common, their variables may be also continuous. In that case, instead of probability tables, pdfs are required.

Example. Fig.3.8 shows a simple BN for diagnosis of excessive dropped calls in cellular networks. The possible causes in this example are *Lack of Coverage* (*no/yes*) and *Interference* (*no/yes*). Their symptoms are *Signal Quality* (*low/normal*) and *Signal level* (*low/normal*). For the quantitative modelling, the probability assessments $P(Coverage)$, $P(Interference)$, $P(Level\,|Coverage)$, $P(Quality|Coverage, Interference)$ are required (Table 3.1).

The chain rule for BNs

Let $U = \{A_1, ..., A_n\}$ be an universe of variables. If the joint probability table $P(U) = P(A_1, ..., A_n)$ is known, then the probability $P(A_i)$ as well as $P(A_i|e)$, where e is evidence, can be calculated from $P(U)$. However, $P(U)$ grows exponentially with the number of variables. Hence, even with a reduced number of variables, the table for $P(U)$ becomes intractably large. Therefore, a method of storing information from which $P(U)$ can be calculated is required. A BN over U is such a representation. If the conditional independencies in the BN hold for U, then $P(U)$ can be calculated from the probabilities specified in the BN. The *chain rule* helps to face that task:

Chain rule: Let B_N be a BN over $U = \{A_1, ..., A_n\}$. The joint probability distribution $P(U)$ is the product of all probabilities specified in B_N:

$$P(U) = \prod_{i=1}^{n} p(A_i|\, \pi(A_i)) \tag{3.3}$$

[10] $P(A = a|B = b)$ is the probability of variable A being a given that variable B is in state b. $P(A|B)$ is a probability table which consists of all the possible combinations of values of A and B

Table 3.1: Probability tables for example in Fig.3.8

$Coverage$=no	0.6
$Coverage$=yes	0.4

$P(Coverage)$

$Interference$=no	0.7
$Interference$=yes	0.3

$P(Interference)$

	$Coverage$=no	$Coverage$=yes
$Level$=low	0.1	0.8
$Level$=$normal$	0.9	0.2

$P(Level|Coverage)$

	$Coverage$=no		$Coverage$=yes	
	$Interference$=no	$Interference$=yes	$Interference$=no	$Interference$=yes
$Quality$=low	0.05	0.7	0.8	0.9
$Quality$=$normal$	0.95	0.3	0.2	0.1

$P(Quality|Coverage, Interference)$

where $\pi(A_i)$ is the parent set of A_i.

Probability updating in BNs can be performed by using the chain rule to calculate $P(U)$. Alternatively, the calculations can be performed without having to deal with the full joint probability table. Additionally, in most cases, Bayes'rule[11] must be also applied.

A variable is *marginalized* out of a probability table if the dependency of the table with that variable is eliminated. In order to marginalize, the sum of the probability table for all possible values of the variable to be marginalized is calculated. For example, if B is marginalized out of $P(A, B)$, the resulting probability table is $P(A)$ and can be obtained as:

$$P(A) = \sum_B P(A, B) \tag{3.4}$$

Evidence propagation is out of the scope of this thesis. Interested readers are addressed to the numerous algorithms found in the literature for belief updating [54, 110, 157, 64].

Example. Consider the BN in Fig.3.8. Assume that we have the evidence $e = \{L{=}normal, Q{=}low\}$ and we wish to calculate $P(C|L{=}normal,\ Q{=}low)$, where I=*Interference*, C=*Coverage*, L=*Level*, Q=*Quality*. In this case:

$$\begin{aligned} P(U, e) &= P(I, C, L = normal, Q = low) = \\ &= P(I)P(C)P(L = normal|C)P(Q = low|I, C) \end{aligned}$$

It should be noticed that calculating the full table $P(U)$ with 2^4 entries is not required. On the contrary, in order to calculate $P(C, e)$ the variable I is marginalized out of $P(U)$:

$$\begin{aligned} P(C, L = normal, Q = low) &= \sum_I P(I, C, L = normal, Q = low) = \\ &= \sum_I P(I)P(C)P(L = normal|C)P(Q = low|I, C) = \\ &= P(C)P(L = normal|C) \sum_I P(I)P(Q = low|I, C) \end{aligned}$$

Subsequently, the two tables $P(I)$ and $P(Q = low|I, C)$ are multiplied and I is marginalized out of the product. This table depends on three variables, instead of the four original ones. In order to calculate $P(C|e)$ we simply apply the Bayes'rule:

$$\begin{aligned} P(C|L = normal, Q = low) &= \frac{P(C, L = normal, Q = low)}{P(L = normal, Q = low)} = \\ &= \frac{P(C, L = normal, Q = low)}{\sum_C P(C, L = normal, Q = low)} \end{aligned}$$

[11] Bayes rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

According to tables 3.1, the probability of having a lack of coverage is 0.33 in this case.

3.2.2 Bayesian modelling

The construction of a model based on BNs can be divided into two parts. The first stage defines the *qualitative model*, which consists of the variables and their relationships. The second part specifies the *quantitative model*, that is the probabilities required for the probability tables in the BN.

The purpose of a BN model is to give estimates of certainties, i.e. probabilities, for events that are not observable (or only observable at an unacceptable cost). These events are called *hypothesis events*. *Hypothesis variables* are groups of mutually exclusive events. In order to obtain a certainty estimate, there should be some information channels which may reveal something about the hypothesis variables. These types of information are grouped into *information variables*.

Having identified the variables for the model, the next step is establishing the directed links. It is essential that the conditional independencies coded in the model correspond with reality.

The quantitative part of the model can be obtained from different sources: subjective probabilities elicited by experts, statistical data or theoretical considerations. Several techniques have been proposed to ease the definition of parameters [175, 188, 76].

Modelling can be simplified if some conditions are assumed about the dependencies in the network. This gives rise to different network structures, such as the *Simple Bayes Model* or simplifications of the required probabilities, such as *Independence of Causal Influence models* [97, 180, 95], which will be both discussed in Chapter 5.

3.2.3 Learning

In some application domains large databases of cases[12] exist, which can be used to build the model. *Learning* is the process of building a BN based on previous training cases. In *structural learning* the network structure that best fits the information is calculated [140, 94, 179, 63, 64, 143]. In *parameter learning* a network structure is assumed and the probabilities are computed based on the cases [71, 49, 127, 83]. Structural learning is out of the scope of this thesis, whereas parameter learning will be dealt in Chapter 5.

3.2.4 Sensitivity analysis

Sensitivity analysis refers to the study of how sensitive the conclusions (probabilities of the hypothesis variables) are to minor changes. The changes may be variations of the parameters of the model or may be changes of the evidence.

For example, for the network in Fig.3.8, a parameter could be the probability $t = P(\mathit{Quality}{=}\mathit{low}|\ \mathit{Interference}{=}\mathit{no},\ \mathit{Coverage}{=}\mathit{yes})$. The expression $P(\mathit{Interference}{=}\mathit{yes}|e(t))$ is the variation of the probability of interference given the evidence vs. the parameter t.

[12] A *case* is the set of values assigned to some information variables in the BN.

One-way sensitivity analysis studies the influence of a single parameter on the conclusions. *Two-way sensitivity analysis* analyzes the impact of changing two parameters simultaneously on the hypothesis variables.

Basically, there are two approaches to sensitivity analysis: theoretical and empirical. The theoretical approach expresses the posterior probability of interest in terms of the parameters under study [110, 55, 126, 120, 81]. The empirical methods examine the effects of varying the parameters of the network on the diagnosis. In the latter case, the most frequent approach consists of adding random noise to the probabilities and examining the effects on the diagnostic performance [103, 118, 158].

Part II

Modelling of Fault Diagnosis

Chapter 4

Diagnosis in GSM/GPRS networks

This chapter is devoted to the description of the knowledge base, i.e. the knowledge about the application domain. Thus, fault diagnosis in GERAN is explained. In order to build a diagnosis model the possible causes which may give rise to the problem and their corresponding symptoms as well as the conditioning parameters have to be identified. Diagnosis is carried out independently in each cell with problems. The identification of the problematic cells is performed by the Fault Detection subsystem, which is outside the scope of this thesis. Therefore, it has been assumed that the diagnosis system is utilized only on cells with problems.

In Section 4.1 the terminology utilized in this chapter is described and the methodological approach is summarized. Section 4.2 describes the main problem in GSM networks, i.e. the situation that triggers the requirement of a diagnosis, which is an excessive number of dropped calls. Different figures to measure dropped calls are also reviewed in that section. The main causes that may provoke dropped calls are depicted in Section 4.3. Afterwards, the information available to support the identification of the cause (named *symptoms*), which are network performance indicators and alarms, are explained in Section 4.4. Furthermore, there are other factors (named *conditions*) that have an influence on the behavior of the network, such as some functionalities and configuration parameters, which are described in Section 4.5. The qualitative relationships amongst causes, symptoms and conditions, together with some examples from a live network, are presented in Section 4.6. Finally, in Appendix 4.A, some specially complex TS cases solved by experts are explained.

Although this chapter is entitled "diagnosis in GSM/GERAN", actually the focus is on GSM, due to the fact that the analysed problem is the excessive dropped call rate. Therefore, a basic knowledge of GSM is required to follow this chapter (see Chapter 2 or [142, 162]).

4.1 Introduction

4.1.1 Basic definitions

A *problem* is defined as a situation in a cell which has a degrading impact on the service offered by that cell. Every operator uses a different method to identify the problematic cells, which can be based on different performance indicators, e.g. dropped calls, access failures, congestion, etc. [123]. The most severe problem for mobile network operators are cells experiencing a high number of dropped calls because a dropped call has a very negative impact on the service offered to the end user. The metrics used to identify these cells varies from operator to operator. Some operators use the absolute number of drops as the determining criteria, others favour the dropped call rate while sometimes the time between two successive dropped calls (Minutes per drop) is also applied to rank cells. Once the cells with problems are isolated, a diagnosis of the cause of the problems should be done for each problematic cell separately.

A *cause* or *fault* is the defective behavior of some logical or physical component in the cell that provokes failures and, finally, generates a problem. An example of physical cause is a fault in a TRX, whereas an example of logical cause is interference in a cell due to a sub-optimal frequency plan.

A *symptom*[1] is a performance indicator or an alarm whose value can be a manifestation of a fault, e.g. the received signal level in the uplink path.

A *failure* is an anomalous value of a symptom, which can be caused by a fault, e.g. excessive number of interference handovers. Therefore, a problem is a type of failure that has a negative influence on the service offered to subscribers.

A *condition* is a factor whose value makes the probability of certain cause occurring increase or decrease. For example, a condition is the functionality of Frequency Hopping (FH)[2] because if it is active the cause interference is less probable than if FH is not activated.

Fig.4.1 depicts the relations amongst the concepts explained above. A cell is at a given state, which may be either correct operation or malfunctioning. The Fault Detection subsystem is in charge of determining which one of the states the cell is in. The values of the symptoms are manifestations of the state of the cell. If the cell is experiencing problems due to a fault, failures come up, i.e. some symptoms display values quite different from those normally seen on the cell. In addition, conditions also impact on the behavior of the cell, sometimes preventing the presence of some faults. Finally, if failures are serious they can affect the client and, therefore, degenerate in a problem. The diagnosis system aims at identifying the cause of the problem.

The complete mobile system includes millions of parameters for all the cells in the network. Ideally all of these have to be taken into account to explain the behavior of the network. However, symptoms and conditions are so numerous that it is not feasible to consider all of them in order to carry out the diagnosis. Furthermore, some issues, such as the geography of the cell or the

[1]At this point, it is important to clarify the use that the word *symptom* will have in this thesis: it does not have any negative connotation, i.e. a symptom can take any value and not only anomalous values. Confusion often arises from the use of the term given in medicine, where a symptom is defined as "a change in the body's condition that is a sign of illness" [106]. A more general definition of symptom is "a sign of the existence of something". For example, fever (temperature>37°C) is a sign of flue. However, the word symptom can be used in a wider sense as a sign of the existence or nonexistence of something. Thus, in the previous example the symptom would be the temperature and not only the fever (which is a range of the temperature) because if temperature is lower than 37°C it is a sign that probably the patient does not have flue. In this thesis, the symptoms will be the alarms and performance indicators, and not only some values of performance indicators (i.e. failures) and active alarms. For example, the temperature or the signal level will be denoted as symptoms, instead of the fever or a signal level lower than X dBm, respectively.

[2]Frequency Hopping is the repeated switching of frequencies during radio transmission.

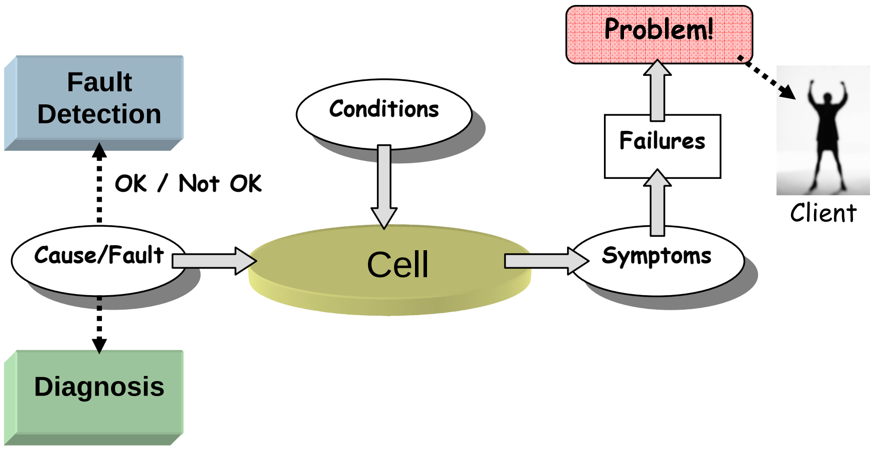


Figure 4.1: Diagnosis concepts

propagation conditions, also have an impact on the cell state. The behavior of a given cell may be even affected by the performance of its neighboring cells. Because of all these reasons, the knowledge base is a model which highly simplifies the behavior of the actual network. The presence of random processes, such as noise in the transmission or the influence of weather on the propagation conditions, together with the impossibility to exactly describe the scenario requires the adoption of non-determinist models.

The aim of the diagnosis system is to identify the cause of a problem based on the values of the symptoms and conditions. Therefore, the problems, causes, symptoms and conditions to be considered in the model should be determined in order to define the knowledge base used by the diagnosis system.

4.1.2 Methodology

The knowledge base presented in this chapter is composed of the causes, symptoms and conditions that operators most frequently use for diagnosis. The information described hereafter is based on interviews with troubleshooting experts of diverse operators and on my experience about radio access networks acquired during the development of this thesis. Thus, the qualitative relations among causes, symptoms and conditions have been defined based on knowledge.

4.2 Problem: Dropped calls

This section is devoted to the definition of the "problem" concept, which will determine under which circumstances diagnosis is required. Indeed, it is assumed that the diagnosis system is used only on cells which are known to have problems. The task of determining whether a cell is experiencing problems is not part of the diagnosis system, but it is assigned to the Fault Detection System (see Fig.2.7).

Each operator uses different metrics to identify which cells have problems. This is due to the fact that no standards exist for fault detection and operators have often slightly different objectives. However, some indicators are common to all operators. Amongst the most important indicators, the high number of dropped calls, the blocked call rate and the high number of access failures stand out. Furthermore, there is no globally accepted measure for calculating them (i.e. the final equation), which can be done in many ways.

This thesis focuses on one of the most important existing problems in GSM/GPRS networks: the occurrence of a high number of dropped calls. Before explaining how to measure the dropped calls, the differences between call and connection should be clarified [162]. Obtaining a traffic channel in a cell is denoted *connection*. A *call* is obtaining a traffic channel in a cell due to a call originated in the mobile phone or finished in the mobile phone. Therefore, the connections of a cell A are composed of the calls done in the cell (initiated or terminated in the MS connected to that cell), the handovers from other cells to cell A and the changes of traffic channels inside cell A (intracell handovers). Thus, for each call there is at least one connection.

A call is *dropped* when it finishes abnormally, i.e. it does not conclude because of the will of one of the two end users. A dropped call has a very negative impact on the quality of service perceived by the end user. This is the reason why achieving a low number of dropped calls is a crucial objective of mobile network operators. Dropped calls are registered in the network statistics and, depending on the manufacturer of the equipment, dropped calls can be even classified according to the reason that caused that drop. It should be pointed out that the aim of operators is not arriving at a 0% of dropped calls because the cost would be too high, but achieving an "acceptable" percentage of dropped calls. The following sentence from an operator summerises the previous philosophy: "we do not want to be *the best*, but be *close* to the best or slightly better than the local competitor". In mature networks, the percentage of dropped calls is normally around 1 or 2% [90].

There are many different formulas to calculate the dropped calls. The most common ones are in one of the following groups:

- *Dropped Call Rate* (DCR): percentage of dropped calls over total number of calls. It is the most common formula to measure dropped calls.

- *Drop-out Ratio*: percentage of dropped calls over total number of connections.

- *Minutes Per Dropped Calls* (MPD): time between two successive dropped calls in a cell. It is a very important indicator because the lower the time between drops the worse the problem. Thus, it is a very efficient metric to identify the worst performing cells.

Previous formulas can vary slightly. For example, some operators consider also call re-establishments in the total number of calls.

Apart from those measurements, it is also important to take into account the absolute number of dropped calls because a high percentage of dropped calls is not so serious in a cell with little traffic, but very worrying in a cell with a significant amount of traffic.

Table 4.1: Dropped call causes

Interference	Uplink Interference Downlink Interference
Coverage	Lack of coverage: borders Lack of coverage: shadows Excessive distance
Hardware	TRXn: Baseband Unit TRXn: general fault Combiner Antenna HW fault in transmission path (connectors, antenna feeder, etc.) HW fault in reception path (MHA, preselector, reception matrices, etc.) Other HW fault
Transmission	A-bis interface fault A interface fault Transcoder
Others	Fading Adjacency definition Configuration parameters

4.3 Causes of dropped calls

Table 4.1 summarizes the most common faults that may cause a high number of dropped calls in GERAN, which will be described in the following sections. Each cause can be split in several subcauses, which could be also considered in the model, e.g. the cause "interference" could be caused by an incorrect frequency plan or by external sources such as a television station. For simplicity sake, in this thesis, only high level causes have been included into the diagnosis model.

4.3.1 Interference

Most current cellular networks are interference limited, due to the tight reuse patterns[3] used. That means that in those networks interference imposes the capacity limit that the network is able to cope with.

Interference causes a high error rate and frame loss, which eventually leads to the release of the radio channel, i.e. to a dropped call. The sources of interference are numerous, the ones that stand out are the following [11]:

- Other mobile network operator is transmitting on the same frequency. Normally this is due to an incorrect setting.
- Cells overlap, i.e. cells from own network exceed the nominal coverage. It can be caused by an insufficient antenna tilt, excessive transmitted power, inadequate frequency plan, a change in the environment (for example, someone may have cut down a forest of trees that had been blocking the signal from that site), etc.

[3]Tight reuse pattern means that the same frequency is used in close cells. Thus, the consequence is that those cells interfere with each other.

- Close channels interference. As the spectrum is becoming saturated, new services are assigned closer frequencies. This may cause interference in the adjacent channels.

- Intermodulation. External signals may cause intermodulation products when they are mixed in the non-linear amplifier of the base-station transmitter. Even physical structures close to the antenna, such as a roof or a fence, can act as non-linear elements, also generating intermodulation products.

- Interference from external transmitters. Close television or radio stations may have harmonics of considerable energy which interfere with the cellular network.

Interference may be present in the uplink path (affecting the reception of the base station) or in the downlink path (affecting the reception of the mobile station). The statistical characteristics of the interference are different in each path because a mobile receives interference from a limited number of fix locations (the base stations), whereas the base stations are interfered by a potentially large number of moving mobile stations.

Actually, the important parameter is not the absolute interference level, but the ratio between the wanted signal level, the carrier C, and the interfering signal level, I, i.e. the ratio $\frac{C}{I}$. On the one hand, C varies with the propagation fluctuations and with the distance between the MS and the serving base station. On the other hand, I basically depends on the distance to the interfering cells, and therefore, on the reuse pattern. The error probability is related to $\frac{C}{I}$. Thus, the network offers an adequate degree of quality against interference only when $\frac{C}{I}$ is higher than a given threshold, denoted *protection ratio*. For example, the Specs[4] [29] establish that the protection ratio for the co-channel[5] interference is 9 dB. Other issues that have an influence in the $\frac{C}{I}$ distribution are the power control, the discontinuous transmission and the frequency hopping, which will be described in Section 4.5.

4.3.2 Coverage

Coverage faults include two types of causes: lack of coverage and excessive distance between MS and BTS. In any case, the final result is a degradation in quality and, finally, a dropped call either because of the low level or the poor quality.

Lack of coverage

There is lack of coverage if the received signal level is below the (MS or BTS) receivers sensitivity.

The coverage area of a cell can be defined by measuring the signal level received from the base stations. That means knowing the C distribution within a cell. Actually, the situation is more complex and cells overlap, i.e. there are areas where an active MS may be connected to various cells, depending on the direction where the MS came from. Hence, the distribution of C

[4] Specs: GSM specifications.

[5] Co-channel interference is interference coming from co-channel cells, that is, other cells which use the same frequency.

in a cell, and therefore the cell area, is determined by the propagation from the own BTS, the propagation from BTS of neighboring cells and the handovers criteria [142].

The lack of coverage has been classified into two types: lack of coverage in the borders of the cell and shadow zones within a cell. Lack of coverage in the borders usually shows up in rural areas, i.e. in areas with low density of population. In these areas, tele-traffic density is normally low and cells tend to have large coverage areas. Those cells are normally limited by the uplink path because the signal levels transmitted by mobiles located close to the cell borders are not enough for the BTS receiver. It can also happen, although less frequently, that mobiles in the borders do not receive with enough power the transmissions from the BTS.

Lack of coverage can also happen in shadow zones within the cell. In those areas the signal level received by either the MS or the BTS is low, e.g. due to the presence of obstacles. This cause may show up both in rural cells and in urban areas, although it is more common in the latter case.

Excessive distance

In this case, a call is dropped because the distance between the MS and the BTS is longer than 35 km. This cause is related with the method to carry out the synchronization on the radio subsystem in GSM networks [28, 162], which is the following. Due to the fact that the channels in GSM are shared according to a TDMA scheme, collisions may occur in the uplink path if two MSs transmit in the same TS (and the same frequency). For example, this situation may happen when two MSs are planned to use consecutive TS and are situated at different distances from the BTS, the second scheduled MS being at a shorter distance that the first one. As a result of the different propagation times, the arrival of the burst of the first scheduled MS gets delayed with respect to the second one, thereby producing an overlap of the signals in time.

The solution in GSM is that the BTS sends to each MS a *timing advance* parameter (TA) according to the perceived round trip propagation delay BTS-MS-BTS. The MS advances its timing by this amount, with the result that signals from different MSs arriving at the BTS are compensated for propagation delay. This process is called *adaptive frame alignment*. In the downlink path these collision problems do not exist because each MS obtains its bursts from the common frame transmitted by the BTS without affecting the other MSs.

The timing advance[6] is in the range 0 to 63. Thus, the maximum distance between the MS and the BTS is 35 Km, which corresponds to TA=63. In rural areas, antennas normally have a low tilt angle because it is not necessary to sacrifice coverage to limit interference, due to

[6] When the BTS detects an access burst transmission on the RACH, it shall measure the delay of this signal relative to the expected signal from a MS at zero distance under static channel conditions. This delay, the timing advance, shall be rounded to the nearest symbol period and be included in a response from the BTS. The BTS shall thereafter continuously monitor the delay of the normal bursts sent by the MS. If the delay changes by more than one symbol period, the timing advance shall be advanced or delayed 1 and the new value signalled to the MS.

Each unit of TA corresponds to a transmission delay equal to a round-trip symbol duration, which, in distance, is equivalent to about 550 m. For example, when TA= 1 the MS is at a distance from the BTS between 550 and 1100 m. The maximum value of TA is 63 because the access burst transmitted by the MS shall be received by the BTS within a given TS and, on the other hand, access bursts have a guard period of duration 68.25 bits.

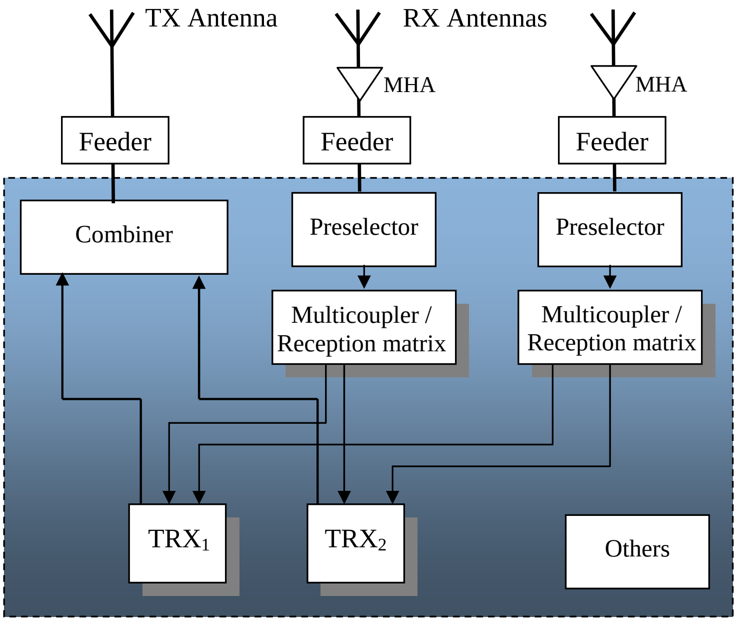


Figure 4.2: BTS Subsystems

the loose reuse pattern. In that case, it may happen that a MS stays connected to a certain BTS even if the distance to another BTS is lower, the reason being that the MS receives higher signal level from the farther BTS than from the closer one. If TA is higher than 63, due to the aforementioned synchronization constraints and no matter what the received power is, the network will automatically drop the call. This problem could be solved by down-tilting the antenna [196], thus forcing calls to be handed off the cell earlier.

4.3.3 Hardware

This cause groups faults of the base station equipment. The base station modules are composed of elements that deteriorate over time, some failing gradually and others suddenly. The effects of hardware faults can go from both reduced signal level and quality to only a quality decrease in one of the radio links. In most cases, when there is a hardware fault numerous alarms are triggered.

Fig.4.2 shows the main functional systems which compose a BTS. The following faults, which corresponds to the diverse subsystems, have been distinguished:

- Fault in one of the TRXs (TRXn). The *transceivers* (Transmitter/Receiver) or *TRXs* are the equipment that manage each of the carriers. The BTSs may have one or more TRXs. A TRX includes a power amplifier for the downlink, a transmitter, a receiver and a baseband unit (which is in charge of coding, interleaving, encryption and assembly in bursts).

- Combiner fault. Combiners are used to connect various transmitters with close frequencies to a single antenna. Isolation among transmitters is assured and the signal from each of them is provide to the antenna with minimum coupling.

- Antenna fault.
- Other faults in the RF transmission chain. It includes faults in the antenna feeder, the connectors, the cables, etc.
- Other faults in the RF reception chain. The reception chain is composed of diverse subsystems: MHA, duplexors, preselectors, antenna multicouplers, switch matrices, etc.
  - The *Masthead Amplifier* (MHA) is a low noise amplifier located between the antenna and the antenna feeder cable. The MHA compensates the cable losses and improves the noise figure of the receiver. The result is an improvement in the uplink chain, a decrease in the link imbalance and, ultimately, an increase of the coverage area.
  - A *duplexer* allows a single antenna to be used to transmit and receive, isolating transmitter and receiver. Typically, duplexers are composed of two bandpass filters, one for the transmitter and the other one for the receiver. In GSM, duplexers are not normally used because individual antennas for transmission and reception are preferred.
  - A *preselector* is composed of a bandpass filter of 25 MHz bandwidth and a RF amplifier. The purpose of the filter is avoiding the entrance of out of band signals, whereas the amplifier provide the required gain to compensate the losses in the antenna feeder, the multicouplers and the reception matrices.
  - The *antenna multicouplers* or *power dividers* have one input and multiple outputs so that multiple receivers can be connected to a single antenna.
  - *Reception matrices* allows the connection of any antenna to any receiver. They include the required power divider modules.
- Other HW faults. Besides the previous elements, base stations include other subsystems, such as synchronization circuits, power supply, A-bis interface connection, air conditioning, etc. Any of these might be the cause of the fault.

4.3.4 Transmission

Transmission causes have been divided into two subcauses: link failure and transcoder failure. The first group includes faults in the A-bis interface and in the A interface.

A-bis Interface faults

These faults can happen at the interface between the BTS and the BSC, i.e. at the A-bis interface [20]. The physical level of this interface is composed of 2 Mbps links (32 channels of 64 kbps), according to the ITU G.703 Recommendation [13].

In the A-bis interface the types of interchanged information are:

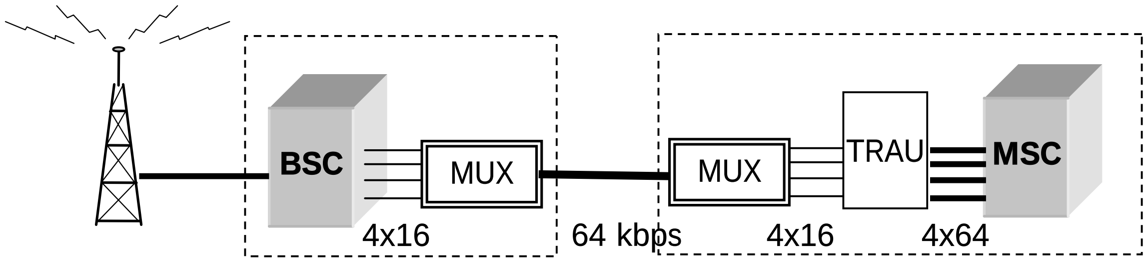


Figure 4.3: TRAU Location

- Messages between the BTS and the BSC in order to control the BTS, by means of the LAPD protocol. The messages are sent to certain piece of equipment (TRX or BCF[7]) so that it carries out a given task.

- Messages between the MS and the BSC or any other part of the network beyond the BSC. These messages transparently cross through the BTS, i.e. they are not analyzed by the BTS.

A Interface faults

Failures of the A-interface [15], i.e. the link between the BSC and the MSC, may also lead to dropped calls. The physical layer of the A-interface is also composed of 2 Mbps digital links, according to G.703 Recommendation.

Transcoder failures

In the air interface, voice channels are coded at 13 kbps, whereas in the rest of the network channels must be coded at 64 kbps, according to the G.703 Recommendation. Therefore, a system to adapt those rates is required: the *Transcoder and Rate Adaptation Unit* (TRAU). Although the TRAU is functionally[8] part of the BTS, normally it is located close to the MSC, as shown in Fig.4.3.

[7]BCF: Base station Control Function. It is the part of the BTS which handles all control functions within the BTS.

[8]The TRAU is functionally considered part of the BTS, although the Specs allow various locations. Firstly, the TRAU could be located in the BSS. In that case, each channel would go at 16 kbps (13 kbps + signalling) between the TRAU and the TRXs, whereas in the A interface transmission would go at 64 kbps. For 4 voice channel, 4 PCM channels at 64 kbps would be required in the A-interface, using only 2 bits per PCM time slot and, therefore, wasting 6 bits.

A more economic location of the TRAU is close to the MSC (Fig.4.3). In that case, in the BSS, 4 channels at 16 kbps could be multiplexed in one single PCM channel. Thus, at the A interface only one link at 64 kbps would be required to transmit the 4 voice channels. In Fig.4.3, it can be observed that between the TRAU and the MSC there are 4 links at 64 kbps because the TRAU only changes the rate, it does not multiplex. If the TRAU and the MSC are in the same site, those links will be local connections.

A BSC can be connected to several TRAUs because each TRAU is used for a 2 Mbps link. A typical TRAU can have as input a link with 120 channels at 16 kbps and an output of 120 channels of 64 kbps.

The most important components of a TRAU are the interfaces with the BSC and the MSC, the system controller and the equipment for transcoding and rate adaptation.

4.3.5 Others

There are other faults not included in the ones explained in the previous sections, which occur more rarely. Among them, the following ones stand out: fading, bad adjacency definition and erroneous configuration parameters.

Fading

Transmission through the mobile channel is subject to fading due to multipath propagation. In channels with multiple reflections, such as in urban areas with several buildings, if transmission is always done on the same frequency it may occur that successive frames might not be decoded. The mobiles more affected by fading are the stationary or very slow ones. Fading causes a high frame error rate, which may lead to dropped calls.

Adjacency definition

In order to execute a handover, quality and level measurements of neighboring cells should be done. With this aim, the BSC shall send to the MS a list of frequencies corresponding to the adjacent cells that it has to measure (*BA list, BCCH Allocation list*) [27]. Therefore, it is important to keep a well defined and up to date adjacency list, so that the handover is always done to the best possible cell.

Two problems may arise [186]:

- Adjacencies that should not be in the list. In that case, the list includes cells from which the average received signal level is too low. Consequently, calls may drop if, at a certain moment, the signal level coming from one of those cells increases and the MS tries to do a handover to that cell.
- Missing important adjacencies. If cells from which the average received signal level is high are omitted in the list, calls may drop when trying to do a handover to cells with lower signal level.

Configuration parameters

In the RAN there are thousands of parameters per BTS, which should be updated when the network evolves. If any of the important parameters is incorrect, dropped calls may appear. An example of parameters are the ones used for handovers, such as the threshold for level handovers.

4.4 Symptoms: Performance Indicators and Alarms

As described in Section 2.1.5, *performance management* (PM), one of the functional areas of network management, handles the execution of performance measurements. In order to do so,

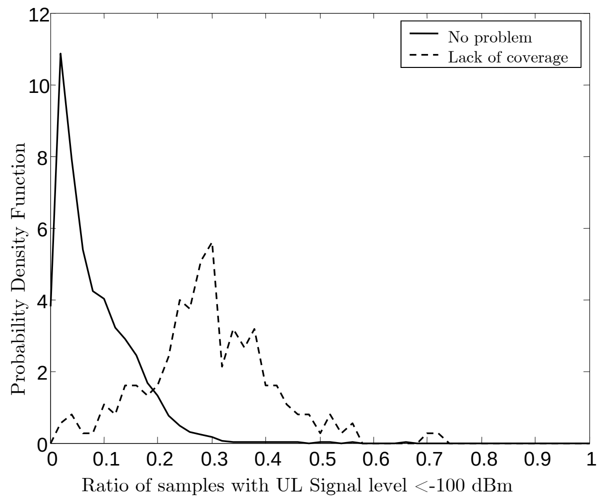


Figure 4.4: Probability density function of symptom depending on fault cause

PM monitors and analyzes the managed network elements [24, 12]. The data which helps to measure the network performance have been called symptoms. Symptoms can be divided into two groups: *performance indicators* and *alarms*.

Amongst performance indicators, the most important ones are called *Key Performance Indicators* (KPIs). KPIs are collected daily by the NMS with the help of counters located in different points of the network [10] (see page 2.1.5). Thus, the NMS contains statistics of thousands of KPIs related to traffic, quality of service, network configuration, failure rate, etc. They can either be referred to the busy hour or be average values for the entire day.

Alarms are the second type of symptoms. Manufacturers of equipment provide hundreds of alarms which are triggered when there is a fault related to a network entity. As this thesis is focused on performance indicators, no deeper information about alarms will be given. Interested readers are referred to the technical documentation of network equipment.

Symptoms are very important for diagnosis because their values can pinpoint certain fault. When the network works as planned, e.g. when the number of dropped calls is low, symptoms display a different statistical distribution compare to the case when a fault is present. Likewise, the distribution of the symptoms are different for each of the causes. For example, the average received signal level from MSs in a cell with a lack of coverage is lower than the signal level in a cell that is operating normally. Fig.4.4 shows the pdf for the ratio of samples with signal level lower than -100 dBm for cells without problems and cells suffering from lack of coverage. In this example, pdfs have been created using histograms from a live network.

Table 4.2 summarizes the main symptoms used for diagnosis, which will be described in the following sections.

Table 4.2: Symptoms

Dropped calls	Number of dropped calls Drop Call Rate (DCR) Radio Failure Component Radio Failure in HO Component A-bis Failure Component A-bis Failure in HO Component LAPD Failure Component A Failure Component A Failure in HO Component Transcoder Failure Component Transcoder Failure in HO Component
Level and quality	RXLEV RXQUAL
Timing Advance	TA
Interference	Interference band
Handovers	Handover Failures Intracell HOs UL/DL Level HOs UL/DL Quality HOs UL/DL Interference HOs Distance HOs Power budget
Access	RACH load
Alarms	see Table 4.6

4.4.1 Symptoms related to dropped calls

Number of dropped calls and DCR

As described in Section 4.2, the main indicator of problems in the network is the presence of a high number of dropped calls[9], which has a very negative impact on end users. There are many different formulas to calculate the dropped calls and each operator uses his own equations. Furthermore, two different symptoms should be separately considered: the absolute number of dropped calls (which is especially important for fault detection) and the percentage of dropped calls (DCR).

Dropped call components

Amongst their KPIs, many manufacturers of equipment provide an estimation of the type of fault that is causing the dropped calls. In most cases, those KPIs do not pinpoint the ultimate fault which is causing the problem, but diverse reasons can increase related counters. For example, the counter related to signalling faults in the A-bis interface can increase if the BSC do not receive a channel activation acknowledgement from the BTS, if measurement reports are not received from the BTS, etc. The ultimate causes of the previous failures can go from a fault in the BTS to interference in one of the TRXs.

KPIs related to dropped call components are normally measured as the percentage of dropped calls due to a given reason. The dropped calls components that have been considered as symptoms for the diagnosis model are the following ones:

- **Radio Failure Component.** In GSM the main immediate cause for the network to drop a call is a *Radio Link Failure*[10]. The reason being that a call should be released by the network when its quality has degraded to a quality under which most clients would manually release the call.

  A radio failure can be caused by diverse faults, which make either the signal level or signal quality be insufficient to sustain the call. Those faults can go from interference or low coverage to problems in the equipment (e.g. water in feeder, lost connections, antenna misalignment, etc.), the environment (e.g. city with many skyscrapers and, therefore, many reflections), etc.

- **Radio Failure in HO Component.** It is similar to the previous fault, but it occurs during a HO attempt. In this situation, a MS has received a *Handover Command*, but

[9] The number/percentage of dropped calls is a *symptom*, whereas the high number/percentage of dropped calls is the *problem*.

[10] In the MS, the procedure to determine that a radio link failure has occurred uses a counter S [27, 26]. If the MS is unable to decode a SACCH message, S is decreased by 1, whereas if a SACCH message is successfully received, S in increased by 2. If S reaches zero, a radio link failure is declared. In any case, S should never exceed the value of *Radio Link Timeout*, which is a value transmitted by each BTS. When there is a radio failure, the MS releases all signalling channels and deactivates the dedicated channels.

On the contrary, the criterion to determine a radio link failure in the BSS is not defined in the Specs and it has to be decided by the operator. Normally, it is based either on the error probability on the SACCH in the uplink or on level and quality measurements.

there has been a failure when trying to connect to the new channels [26]. In that case, the MS would deactivate the new channels and reactivate the old TCH channel and associated signalling channels. After that, a *Handover Failure* message would be sent to the BTS and the MS would resume normal operation as if the HO attempt had not occurred. However, if a failure occurs when trying to reactivate the old channels the call will drop. If the reason of the unsuccessful reconnection has been a radio failure, the KPI "Radio Failure in HO Component" will increase.

- **A-bis interface Failure Component.** The counter associated to this component increases when a signalling message [22] from BTS to BSC is lost (e.g. missing acknowledgement of channel activation, missing establishment indication, missing handover detection, etc.), when an error indication is received from the BTS (because of a protocol error in the air interface), when measurement reports are no longer received from BTS, when channel assignment is not completed because no response is received from the MS before a timer elapses [26], etc. Thus, if this counter increases, it can be due to diverse reasons, apart from a fault in the physical connection, going from a faulty BTS to interference in a TRX.

- **A-bis interface Failure in HO Component.** This KPI indicates the same fault as the previous one, but it increases when trying to return to the old channel after an unsuccessful HO.

- **LAPD Failure Component.** This component increases when LAPD failures take place. As explained in section 4.3.4, LAPD is the link layer protocol used for signalling on the A-bis interface [21].

- **A-interface Failure Component.** This component is related with the protocols used in the A interface [15, 19, 23]. For example, the counter associated to this component may increase when a *Clear Command* is received from the MSC during the call setup phase before the radio resources are assigned to the MS. Another reason for this indicator to increase is an abnormal clear received due to signalling problems in the A interface.

- **A-interface Failure in HO Component.** It is the same fault as the previous one, but when trying to return to the old channel after an unsuccessful HO.

- **Transcoder Failure Component.** This component indicates a transcoder failure during a call attempt. The associated counter is updated when the BTS sends a *Connection Failure Indication* to the BSC with the cause "remote transcoder failure" [22]. This counter may increase not only when there is a transcoder fault, but also in case of a fault in the BTS.

- **Transcoder Failure in HO Component.** It is the same fault as the previous one, but when trying to return to the old channel after an unsuccessful HO.

4.4.2 Level and quality

Symptoms described hereafter are performance indicators obtained from radio link measurements, which are continuously carried out by the MSs (downlink) and the BTSs (uplink) during a call. These measurements were defined in the Specs to be used by the handover and power control algorithms [27]. Two classes of measurements are performed: level and quality measurements. On the one hand, BTSs measure level and quality in the uplink path. On the other hand, MSs carry out quality and level measurements for the downlink path and they send report messages, called *Measurement Reports*[11], to the BSS with those level and quality measurements.

In any case, those measurements are quantified into intervals and symptoms are normally calculated as the percentage of samples[12], during a measurement period (e.g. day, hour, etc.) per range of values. How level and quality measurements are carried out and how the intervals are defined will be explained hereafter.

Signal level

Both the MSs and the BSSs measure the received signal level at the receiver input over the full range of -110 dBm to -48 dBm. Then, the average (in dBm) of the signal level measured samples within the reporting period of length one SACCH multiframe is calculated. The measured signal level shall be mapped to a parameter, named **RXLEV**, between 0 and 63, according to the following intervals:

$$
\begin{aligned}
&RXLEV = 0 \Leftrightarrow level \leq -110 \;\; [dBm] \\
&RXLEV = n \Leftrightarrow -110 + n - 1 \;\; [dBm] \leq level \leq -109 + n - 1 \;\; [dBm] \;\; ; \; n = 1..62 \\
&RXLEV = 63 \Leftrightarrow level > -48 \;\; [dBm]
\end{aligned}
\tag{4.1}
$$

[11]Quality and level measurements for the downlink path are carried out by the MSs and they are sent to the BSS onto the SACCH channel in report messages called *Measurement Reports*. Measurement Reports contain the following information:

- Level received from the serving cell, RXLEV_FULL and RXLEV_SUB. RXLEV_FULL is calculated as the average of each of the measurements taken in the TCH and SACCH frames of the SACCH multiframe, i.e. the average of $25 \cdot 4 = 100$ measurements of the received signal level in the downlink traffic channel. RXLEV_SUB is calculated when using discontinuous transmission. It is the average of the measurements taken in the SACCH frames of the SACCH multiframe and in the frames used to update the characteristics of the comfort noise of discontinuous transmission, SID frames (see Section 4.5.3). In total, RXLEV_SUB is the average of 12 level measurements.
- Quality received from the serving cell, RXQUAL_FULL and RXQUAL_SUB. RXQUAL_FULL is calculated as the equivalent BER before channel decoding during a SACCH multiframe. RXQUAL_SUB is computed in a similar way, although in this case there are only 12 received frames: 4 SACCH frames and 8 SID frames used for discontinuous transmission.
- Level received from the adjacent cells. The network sends a list (*BA-list*) containing the frequencies that the MS should measure. Then, the MS measures the RXLEV for the BCCH carriers of the cells in the list and ranks them in descending order. The measurement report contains the 6 higher RXLEV, together with their corresponding frequencies and BSIC.

RXLEV and RXQUAL measurements are usually filtered before being used by the network [8]. This is equivalent to an average of the measurements, which softens the noise contribution and decreases the impact of fading.

[12]A *sample* is the value measured by the MS or BSS during a SACCH multiframe period. A SACCH multiframe comprises 102 (470.8 ms) or 104 (480 ms) TDMA frames, for a TCH channel and a SDCCH channel, respectively.

Table 4.3: RXQUAL - BER Mapping

RXQUAL	BER	Assumed value
0	$BER < 0.2\%$	0.14%
1	$0.2\% < BER < 0.4\%$	0.28%
2	$0.4\% < BER < 0.8\%$	0.57%
3	$0.8\% < BER < 1.6\%$	1.13%
4	$1.6\% < BER < 3.2\%$	2.26%
5	$3.2\% < BER < 6.4\%$	4.53%
6	$6.4\% < BER < 12.8\%$	9.05%
7	$12.8\% < BER$	18.10%

Finally, level related symptoms are normally either in the form of average level during the measurement period (e.g. day, hour, etc.) or they provide the percentage of samples in the measurement period per interval or set of intervals. An example of level related symptom is the "percentage of samples in the uplink whose RXLEV<10".

Signal quality

Both the MSs and the BSSs should also measure the received signal quality in a manner that it can be related to an equivalent average BER before channel decoding, assessed over the reporting period of one SACCH multiframe. A parameter, named **RXQUAL**, is defined, which quantifies the BER according to the rule shown in Table 4.3.

Symptoms related to RXQUAL are normally calculated as the percentage of samples in the measurement period per range (RXQUAL=0..7) or group of bands (e.g. share with RXQUAL=5-7). For example, a symptom could be the percentage of uplink quality samples out of band 0, i.e. measurements with RXQUAL in bands 1 to 7 in the uplink path. If this symptom is high, the cell is suffering from bad quality in the uplink.

4.4.3 Timing Advance

As explained in Section 4.3.2, the BSS calculates the delay of the signal received from each MS relative to the expected signal from a MS at zero distance [28]. The distance of the MS to the BTS is related to this delay. The TA parameter quantifies the delay in units of a symbol period (approx. 3.69 μs), which is related to the distance according to:

$$TA = \left\lfloor \frac{distance\,[m]}{550} \right\rfloor \tag{4.2}$$

where $\lfloor x \rfloor$ is the integer part of x.

KPIs related to TA are normally calculated as percentage of samples during the measurement period within a range of TA, e.g. percentage of samples with TA>10.

Table 4.4: Interference Mapping

Interference Band	Limits
1	0 (-110 dBm) $\leq$ Level $\leq X_1$
2	$X_1 \leq$ Level $\leq X_2$
3	$X_2 \leq$ Level $\leq X_3$
4	$X_3 \leq$ Level $\leq X_4$
5	$X_4 \leq$ Level $\leq X_5$

Table 4.5: Example of interference mapping

-110 dBm $\leq$ Band 1 $\leq$ -105 dBm
-105 dBm $\leq$ Band 2 $\leq$ -100 dBm
-100 dBm $\leq$ Band 3 $\leq$ -95 dBm
-95 dBm $\leq$ Band 4 $\leq$ -90 dBm
-90 dBm $\leq$ Band 5 $\leq$ -47 dBm

4.4.4 Measured Uplink interference

The BSS measures the interference level in the uplink on the idle time slots and calculates the average over a certain number of consecutive SACCH multiframes [27]. The result is mapped into one of the interference categories shown in Table 4.4. The limits between bands, X_i, are set by operators and they can vary from cell to cell. For example, some typical values are shown in Table 4.5.

Related symptoms are measured as the percentage of samples within each range or sets of ranges. This symptom pinpoint an interference fault if the percentage of samples out of band 1 is greater than certain threshold, e.g. a typical threshold is 3%.

4.4.5 Handovers

The handover (HO) is a mechanism to transfer a call from cell to cell when a MS is moving. One of the aims of HO is that the MS always stays connected to the best serving cell, performing the change between cells without the client to notice.

The HO procedure [26] is always initiated by the network by sending a *Handover Command* to the MS, which includes the characteristics of the new cell to where the HO is taking place. When the MS receives the message, it initiates the release of the link layer connections, it disconnects the physical channels and it initiates the lower layer connections in the new channel. Once connections in the new channel are ready, the MS sends a *Handover Complete* message to the network to indicate the handover completion. Then, the network releases the old channels.

There are some cases of abnormal HO completion. For example, if in the HO request message the network instructs the MS to use a non-supported frequency or if a lower layer failure occurs. In those cases the MS deactivates the new channels, reactivates the old channels and sends a *Handover Failure* message. As a consequence the network releases the new channels. Therefore, problems in a cell are normally reflected in an increment of the following symptoms:

- **Number of HO failures**. On the one hand, if the downlink path suffers from interference, then quality is degraded and the MS is unable to decode the *Handover Command* message sent by the network. Therefore the HO never happens. On the other hand, if quality is bad in the uplink path of the target cell, the MS will receive the *Handover Command*, but the BTS will not decode the *Handover Complete* messages sent by the MS and, therefore, in this case the HO will also fail.

- **Number of intra-cell HOs**. In bad quality conditions, if the signal level is high enough, normally intra-cell HOs are preferred to inter-cell HOs. Thus, if quality is bad, intracell HOs are frequently attempted.

There are different HO algorithms, but most of them are based on level, quality and timing advance measurements. Symptoms related to HO schemes are normally measured as the percentage of HOs of each of the following types:

- **Level HO.** Level HO takes place when the received signal level (RXLEV) by/from the serving cell is below a threshold set by the operator. This situation is typical when the MS is at the border of the serving cell.

- **Quality HO.** Quality HO occurs because the signal quality (RXQUAL) on the uplink or downlink path is below a threshold set by the operator.

- **Interference HO.** Interference HOs are particular types of quality HOs. Interference HO takes place when level is over a certain threshold, but quality is below another threshold. Among these HOs, the intra-cell HOs consist on the change from a time slot to another one within the same serving cell.

- **Distance HO.** Distance HO takes place when the measured TA is higher than a threshold set by the operator. This type of HO normally occurs in rural areas, in which the MS is at the cell border and it does not receive high signal levels from other cells. Another typical scenarios are seaside areas, where due to the signal reflections onto the sea, the signal can reach large distances and, therefore, distance HOs take place to avoid interference.

- **Power budget HO.** The MS continuously evaluates whether the signal levels received from the adjacent cells are higher than the level received from the serving cell. Power budget HO occurs when the level received from a neighboring BTS is above the level received from the serving cell plus a margin.

HOs algorithms differ from vendor to vendor and the operator can decide which HO types to enable, as well as their parameters and priorities. For example, a given operator could have the following priorities in its HOs: 1) Intra-cell, 2) Interference, 3) Level, 4) Distance, 5) Quality, 6) Power budget.

Table 4.6: Alarms

A interface alarm
A-bis interface alarm
LAPD alarm
Transcoder alarm
Combiner alarm
Antenna alarm
BTS with no transactions alarm
Reduced output signal alarm
Transcoder related alarm
Excessive interference alarm
Path difference alarm
Channel failure rate alarm
HW alarm

4.4.6 Access

As described in 2.1.4, the random access channel RACH is the logical channel used by the MS to transmit a *Channel Request*[13] message when it wants to access the network [26]. Among the symptoms related to RACH accesses, the RACH load stands out, which is the percentage of the total RACH capacity that is being used on average during the measurement period. This indicator normally increases when there are problems caused by interference in the uplink path.

4.4.7 Alarms

Current network management systems provide thousands of alarms, which are triggered by certain events. Alarms are not always an indication of a fault. For example, if network traffic increases, an alarm may be triggered, which would not be a symptom of a fault. Furthermore, a fault in a piece of equipment may affect other pieces of equipment, causing their alarms to be triggered. This is the reason why identifying the cause of a problem based only on alarms is very difficult.

Alarm correlation [109, 192, 202] is the interpretation of multiple alarms so that a new meaning is given to the original alarms (see Chapter 2). Multiple references can be found about alarm correlation for fault diagnosis, whereas no attention has been paid to automatic diagnosis based on performance indicators. This thesis is focused on performance indicators because of the previous reasons. Nevertheless, some alarms have also been included in our knowledge base (Table 4.6), which can be directly obtained from the NMS or after an alarm correlation process.

[13]The number of RACH time slots between initiation of the immediate assignment procedure and the first *Channel Request* is a value drawn randomly with uniform probability distribution for each new initial assignment initiation. After sending the *Channel Request* message, the MS starts listening to the BCCH waiting for the network reply. If it does not receive any response, the MS sends a new *Channel Request*. The number of RACH time slots between two successive *Channel Request* is also a random value from a uniform probability distribution. The maximum number of channel requests that the MS can send is broadcasted in the BCCH. Once the last request is transmitted, a timer starts and when it expires the immediate assignment procedure is aborted and a random access failure is indicated.

Table 4.7: Conditions

Features	Configurations
Frequency Hopping	Cell type
Power Control	Climate
Discontinuous Transmission	Antenna Alignment
Reception Diversity	Antenna tilt

4.5 Conditions: Features and Configurations

In this section, conditions, which are factors having an impact on the network behavior, are described. Conditions have been divided into two groups: features and configurations (Table 4.7). *Features* are functionalities applied by operators in order to improve the quality in the communications, e.g. frequency hopping. *Configurations* are special characteristics of certain cells which determines the faults that may occur in those cells, e.g. cell type (rural or urban cells).

4.5.1 Frequency Hopping

Frequency Hopping (FH) consists in changing the frequency assigned to certain communication channel at each transmission burst. Frequency hopping occurs between time slots and, therefore, a mobile station transmits (or receives) on a fixed frequency during one time slot and then it must hop[14] before the time slot on the next TDMA frame [18].

The purpose of FH is diminishing the negative effects of multipath fading and interference. On the one hand, thanks to FH, if a frame is suffering from fading it is very probable that the following frame will not be affected because it will be transmitted onto a different frequency (*frequency diversity*). On the other hand, in a cell suffering from interference, if FH is activated in the network, the interfering transmitter will be changing its frequencies and, therefore, at

[14] The beacon channel, BCCH, should be transmitted at a constant frequency, without FH. Therefore, one of the transmitters should be reserved for that frequency. For the other transmitters, there are two types of FH: *baseband FH* (BB FH) and *radiofrequency FH* (RF FH). In BB FH each transmitter works at a fix frequency and the successive bursts change the transmitter along the communication. In RF FH, also called *synthesized FH*, each transmitter handles all bursts belonging to the same connection. Therefore, transmitters should change their frequency from burst to burst.

Among the configuration parameters related to FH are the hopping frequencies and the hopping sequence, which is identified by the *Hopping Sequence Number* HSN. The HSN can take values from 0 to 63. There are different types of hopping sequences, the main ones are:

- Cyclic sequences. They are mainly used to diminish the effects of multipath fading. The available frequencies are used in consecutive order and HSN is equal to zero. For example, if the hopping group is composed of 4 frequencies, the cyclic distribution of frequencies will be: $...f_1, f_2, f_3, f_4, f_1, f_2, f_3, f_4, f_1, ...$.
- Random sequences. They are used to diminish the effects of co-channel interference. The available sequences are pseudo-randomly ordered. Specs define 63 different sequences depending on the degree of randomness. For example, a random sequence could be $...f_3, f_2, f_2, f_1, f_1, f_4, f_3, f_1, ...$.
- Orthogonal sequences. All transmitters in a cell have the same frequencies and the same hopping sequences. An offset is applied to the frequencies assigned to each transmitter so that the transmitters do not interfere, which is determined by the parameter *Mobile Allocation Index Offset*, MAIO. Thus, although all the transmitters use the same sequence, channels coinciding in time do not use the same frequency.

each moment the only interference will come from those transmitters whose frequencies will coincide with the frequency of the wanted signal (*interference diversity*). Thus, interference changes in each burst, which benefits the connections. If FH were not used, a connection could suffer interference during all the call duration. Hence, FH converts a continuous interference into interference on isolated bursts, which can be counteracted more easily with interleaving and coding techniques.

FH is mainly used in urban areas. In other scenarios, such as roads, fading is less probable to occur because of two reasons. Firstly, the largest signal contribution comes from the direct ray. Secondly, the channel characteristics change very rapidly, due to the high speed. In addition, in urban areas tight reuse patters are normally used in order to increase capacity. The consequence is that co-channel interference is larger in cities than in rural areas.

4.5.2 Power Control

Power Control (PC) adapts the power transmitted by both the MS and the BTS to the propagation conditions [29, 27]. This feature can be separately activated on the uplink or the downlink, although the most common situation is that both links have activated power control[15]. The aim of PC is minimizing the transmitted power while maintaining the good quality of the radio links.

Power control algorithms are based on the level and quality measurements described in 4.4.2, which are filtered before being used by PC. Power is increased if level or quality are low, and it is decreased in the opposite case.

The following benefits are derived from the use of PC:

- Increase in the MS battery duration.
- Decrease of the saturation risk in the receiver. When a MS is very close to the BTS, if PC was not used the high signal level could saturate either the MS or BTS receiver.
- Interference reduction if all MSs and BTSs in the network use PC.
- Quality improvement, because if a bad quality is measured, the transmitted power is increased.

4.5.3 Discontinuous Transmission

Discontinuous Transmission (DTX) [17] is a mechanism which allows the radio transmitters to be switched off most of the time during speech pauses[16]. DTX has the following two purposes: to save power in the MS and to reduce the overall interference level.

[15]In the BCCH carrier power control is not applied and the BTS should always transmit at the maximum power set for that cell.

[16]A basic problem when using DTX is that the background acoustic noise, which is transmitted together with the speech, would disappear when the radio transmission was cut, resulting in a modulation of the background noise. Since the DTX switching can take place rapidly, it has been found that this effect can be very annoying for the listener. The way to overcome this problem is generating on the receiver side synthetic noise similar to the transmitter side background noise, which is called *confort noise* [16]. The parameters of this comfort noise are estimated on the transmitter and send to the receiver before the radio transmission is cut and at a regular

4.5.4 Reception diversity

One of the most commonly applied techniques to fight against the effects of multipath fading is reception diversity. It is based on receiving several poorly correlated signals, in such a manner that the probability of both signal fading simultaneously is negligible.

Currently, the most frequently used diversity method in base stations is *space diversity*, which utilizes two reception antennas conveniently separated. Another technique to implement diversity is by means of *polarization diversity*, which uses a single receiving antenna with double polarization. The received signal is obtained either by combining the outputs of each reception chain[17] or by selecting one of the outputs.

4.5.5 Cell type

Three types of cells have been distinguished, *low*, *normal* or *dense*, depending on the population density on that cell. Certain characteristics related to each type of cell have an impact on their associated faults. For example, cells whose "Cell Type" is *low* are normally large cells. Furthermore, dropped calls in large cells are usually caused by a lack of coverage. On the contrary, normally the cell size is small for cells whose "Cell Type" is *dense*. In addition, in small size cells failures are commonly caused by interference.

Factors related to the "Cell Type" which affects the network performance are the directivities of the antennas, the reuse factor and the cell size:

- **Antenna directivity.** In order to diminish interference and, therefore, increase network capacity, sectorization is used. Sectorization divides a cell with an omnidirectional antenna in sectors, each sector with a directive antenna.

- **Reuse factor.** Frequency reuse in cellular networks allows to increase the network capacity. Nevertheless, it is required to maintain a distance between co-channel cells due to potential interference problems. The *reuse factor* indicates the number of cells in a *cluster*, i.e. in a group of cells where the frequencies belonging to an operator can only be assigned to one cell. Normally, the reuse factor is expressed as x/y, where x is the reuse factor per base station (i.e. per site) and y is the reuse factor per cell. For example, a reuse factor of 3/9 consists of tri-sectorized sites and each frequency is used only once every three sites, i.e. every 9 cells. In the case of a reuse factor 3/3 each frequency is used only once every 3 sites, which are omnidirectional. For a given cell size, the lower the reuse factor, the lower the co-channel distance, and therefore, the lower the ratio $\frac{C}{I}$ is.

low rate afterwards. This allows the comfort noise to adapt to the changes of the noise on the transmitter side. The comfort noise parameters are encoded into a special frame, called SID (*Silence Descriptor*) frame. When the transmitter detects no speech, it sends a SID frame and transmission stops. Until no speech is detected again, 8 SID frames are sent in each SACCH multiframe. When DTX is activated, during the silence intervals the measurements used for the handover and power control algorithms are the SUB measurements, whereas during conversation FULL measurements are utilized.

[17] Combiners are designed so that input signals are coherently mixed.

Table 4.8: Cell Size

Name	**Radius (Km)**	**Typical scenario**
Large cells or Macrocells	1.5 - 20	Roads and rural areas
Small cells or Minicells	0.7 - 1.5	Rural areas
Microcells	0.3 - 0.7	Cities with dense traffic
Picocells	0.03 - 0.3	Commercial centers, airports, offices, etc.

Table 4.9: Characteristics of rural and urban areas

Rural areas	**Urban areas**
Scarcely populated zones, such as small villages or road	Zones with high traffic density
Noise-limited cellular systems	Interference-limited cellular systems
Large cells	Small cells
Loose reuse factor, due to the fact that there are no capacity limitations	Tight reuse factor (in order to increase network capacity)
Sites located in dominant places, such as top of mountains	Sites located below roof top level, but in places without close obstacles
Omnidirectional antennas	Tri-sectorized cells with directional antennas
Low tilted antennas	Quite downtilted antennas
	Reception diversity is used

- **Cell size.** Reduction of the cell size is a method often used to increase the network capacity. In any network, different cell sizes coexist: small cells in areas with high traffic density and large cells in areas with low traffic density. Depending on their size, cells can be classified into the groups shown in Table 4.8 [162, 25].

Two completely different scenarios can be distinguished, rural or urban areas, which determines the network design, and therefore, the value of the three parameters described above. Typical characteristics of urban and rural areas are presented in Table 4.9. For cells located in rural areas, "Cell Type" will be *low*, whereas for cells in highly populated urban areas, "Cell Type" will be *dense*. "Cell Type" will be *normal* for cells in areas whose characteristics are in the middle of the other two types.

4.5.6 Climate

The climate may have an impact on the faults occurring in the cells. For example, in moist climates where rain is frequent, the probability of hardware faults increases. This is due to the fact that water can easily get into a piece of equipment or some connections may come loose. This also has an impact on the quality of transmission links, if microwave radio are used in the A-bis links. The Configuration denoted as "Climate" may take on the values *wet*, *normal* or *dry*.

4.5.7 Antenna alignment and tilt

"Antenna alignment" has a direct impact on the network performance. If the antenna is misaligned dropped calls may happen because of two main reasons. On the one hand, the signal level may be too low in some areas, causing dropped calls in the serving cell. On the other hand, interference from the cell may cause dropped calls in other cells.

Another configuration parameter is the "antenna tilt". If the antenna is downtilted too much, dropped calls may show up because of lack of coverage in the border areas. If the tilt is not enough, dropped calls may appear in the serving cell because of excessive timing advance and also interference may cause dropped calls in other cells.

4.6 Knowledge base for diagnosis model

In this section, the knowledge base for diagnosis in the RAN of GSM networks is summarized. The knowledge base is composed of three types of elements: causes, symptoms and conditions. Table 4.1 showed the causes considered for the model, whereas conditions were enumerated in Table 4.7. In order to simplify the model, subcauses have not been differentiated. For example, the cause "UL interference" can be caused by several reasons (wrong frequency plan, intermodulation, etc.). A more complex model should distinguish among those faults.

In the case of the symptoms, KPIs have been calculated as percentages of samples (=measurements) complying with a given condition during the measurement period (e.g. a day, busy hour, etc.). An example of symptom is the percentage of samples with RXQUAL=1 to 7. If this KPI is high, it indicates bad quality. This approach follows the method used by network operators to calculate KPIs used for diagnosis. Table 4.10 shows the symptoms considered in the model, together with the code and the number per symptom used along this thesis.

Once the elements are defined, the following step is specifying the relationships among them. Firstly, conditions have an impact on certain causes. Secondly, each cause is related to some symptoms. Table 4.11 presents conditions and symptoms related to each of the causes. Symptoms have been defined so that their values increase when any of the related causes is present. It should be pointed out that the symptoms for each cause in Table 4.11 are the most important ones. However, other symptoms may also vary their value when the cause occur. For example, although the lack of coverage decreases the quality, symptoms related to RXQUAL are not shown in the table because the impact of coverage on other symptoms is much higher. The following section presents some examples of causes and the corresponding values of the symptoms in a live GSM/GPRS network.

4.6.1 Example of symptom values

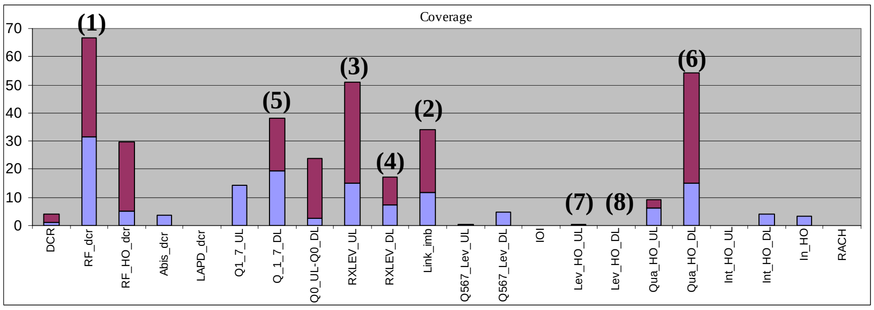
In order to illustrate the concepts explained in the present chapter, in this section some real cases from a live GSM/GPRS network are presented. In Fig.4.5, the values of some symptoms are shown for four faulty cells. The heading of each graphic is the cause diagnosed by experts in troubleshooting. Parts of the bars have been drawn in dark if a failure occurs (i.e. the value for that symptom is abnormal, according to thresholds defined by experts).

The first example in Fig.4.5(a) is a cell suffering from lack of coverage. It can be noticed that the percentage of dropped calls due to radio failures is high (1). The received signal level is lower in the uplink than in the downlink (2), although in both links an appreciable reduction in the received power is observed (3-4). Quality is specially bad in the DL (5) and, therefore, DL quality HOs are also high (6). Nevertheless, we would expect level HOs also to be high, but they are practically nonexistent (7-8). The reason may be that level HOs are very low in the HO ranking of priorities set by the network operator. In that case, level HOs will never happen because a quality HO will probably be triggered beforehand.

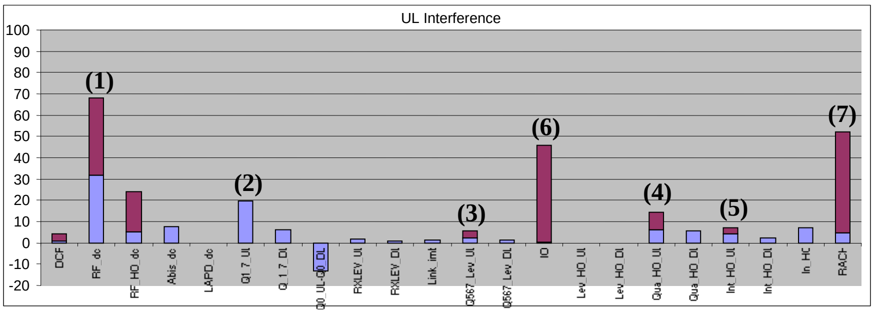
The cell in the second example (Fig.4.5(b)) suffered from interference in the uplink path. The percentage of dropped calls due to radio failures is high (1), as expected. The percentage of samples with bad quality in the uplink path is also high, although not excessive to be considered serious (2). In addition, the number of samples in the UL suffering from bad quality whose level is high is considerable (3). Another consequence of the interference problem is that the number of HOs due to bad quality and interference in the uplink rises (4-5). It can also be appreciated that the interference measured on the idle slots in the uplink is quite high (6). The RACH load also goes up due to the interference (7).

The third example (Fig.4.5(c)) presents symptoms for a cell with a HW fault in the transmission part. The dropped calls due to radio failures are high (1). Quality in the DL path is bad (2) and worse than quality in the UL path (3). The share of quality HOs, especially DL quality HOs (4), is high. Signal levels both in the UL and DL has gone down (5-6). Reduced level in the UL may indicate that diagnosis asserted by experts was wrong because the fault was not only in the transmission path, but common to transmitter and receiver. It may be possible that the problem was due to a faulty TRX, which would also explained the increased number of LAPD failures (7). Alarms should be examined in order to achieve a conclusive diagnostic.

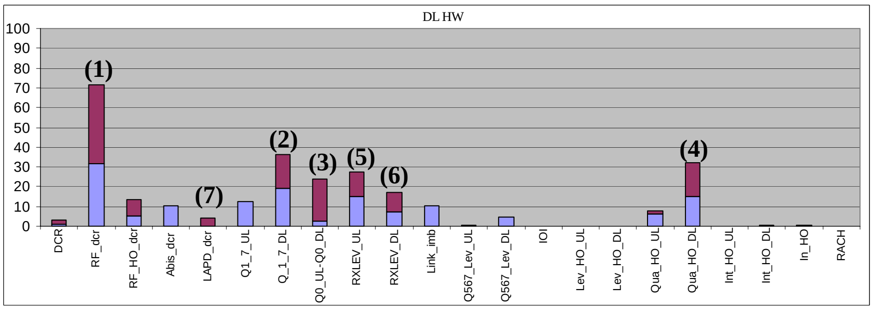
The last example (Fig.4.5(d)) presents the symptoms for a cell with a fault in the A-bis interface. The main symptom is the indicator of the increased number of dropped calls due to A-bis failures (1).



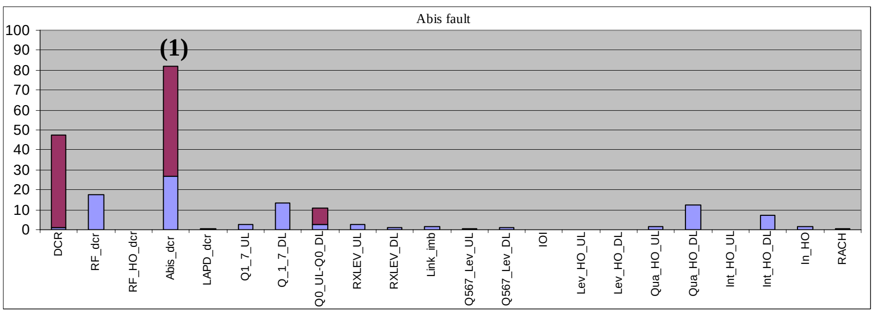
(a)



(b)



(c)



(d)

Figure 4.5: Values of performance indicators for some cases

Table 4.10: Symptoms in the diagnosis model

Number	Code	Name	Description
		Dropped Calls	
1	DCR	Drop Call Rate	Dropped calls / Total number of calls
2	RF_dcr	Dropped calls caused by radio failure	Dropped calls due to radio failure / Total number of calls
3	RF_HO_dcr	Dropped calls caused by radio failure in old channel during HO	Dropped calls due to radio failure after HO failure / Total number of calls
4	A-bis_dcr	Dropped calls caused by failure in A-bis interface	Dropped calls due to failure in A-bis interface/ Total number of calls
5	A-bis_HO_dcr	Dropped calls caused by failure in A-bis interface in old channel during HO	Dropped calls due to failure in A-bis interface after HO failure / Total number of calls
6	A_dcr	Dropped calls caused by failure in A interface	Dropped calls due to failure in A interface / Total number of calls
7	A_HO_dcr	Dropped calls caused by failure in A interface in old channel during HO	Dropped calls due to failure in A interface after HO failure / Total number of calls
8	Tr_dcr	Dropped calls caused by transcoder failure	Dropped calls due to transcoder failure / Total number of calls
9	Tr_HO_dcr	Dropped calls caused by transcoder failure in old channel during HO	Dropped calls due to transcoder failure after HO failure / Total number of calls
10	LAPD_dcr	Dropped calls caused by LAPD failure	Dropped calls due to LAPD failure / Total number of calls
		Level and Quality	
11	Q1_7_UL	Percentage of UL samples with RXQUAL out of band 0	Samples with UL RXQUAL=1-7 / Total number of UL samples
12	Q1_7_DL	Percentage of DL samples with RXQUAL out of band 0	Samples with DL RXQUAL=1-7 / Total number of DL samples
13	Q0_UL-Q0_DL	DL quality link imbalance	(Samples with UL RXQUAL=0 / Total number of UL samples)-(Samples with DL RXQUAL=0 / Total number of DL samples)
14	Q0_DL-Q0_UL	UL quality link imbalance	(Samples with DL RXQUAL=0 / Total number of DL samples)-(Samples with UL RXQUAL=0 / Total number of UL samples)
15	RXLEV_UL	Percentage of UL samples with level< -100 dBm	Samples with UL RXLEV< 10 / Total number of UL samples
16	RXLEV_DL	Percentage of DL samples with level< -100 dBm	Samples with DL RXLEV< 10 / Total number of DL samples
17	Link_imb	Link imbalance	abs(RXLEV_DL-RXLEV_UL)

Number	Code	Name	Description
18	Q567_Lev_UL	Percentage of UL samples with RXQUAL in bands 5,6 and 7 with level > -95 dBm	Samples with UL RXQUAL=5-7 and UL RXLEV>15/ Total number of UL samples
19	Q567_Lev_DL	Percentage of DL samples with RXQUAL in bands 5,6 and 7 with level > -95 dBm	Samples with DL RXQUAL=5-7 and DL RXLEV>15/ Total number of DL samples
20.n	Q5X6X7_TRXn	Percentage of samples with RXQUAL in bands 5,6 or 7 in TRXn	max(Samples with RXQUAL=5, 6 or 7 in TRXn)/ Total number of samples in TRXn
Timing Advance			
21	TA_10	Percentage of samples with TA>10 (5.5 Km)	Samples with TA>10 / Total samples
22	TA_4	Percentage of samples with TA<4 (2.2 Km)	Samples with TA<4 / Total samples
23	TA_50	Percentage of samples with TA>50 (27.5 Km)	Samples with TA>50 / Total samples
Interference			
24	IOI	Percentage of samples on UL idle channels out of band 1	Samples in bands 2-5 / Total samples
Handovers			
25	HO_F	Percentage of HO failures	HO failures / Total number of HOs
26	Lev_HO_UL	Percentage of UL Level HOs	UL Level HOs / Total number of HOs
27	Lev_HO_DL	Percentage of DL Level HOs	DL Level HOs / Total number of HOs
28	Qua_HO_UL	Percentage of UL Quality HOs	UL Quality HOs / Total number of HOs
29	Qua_HO_DL	Percentage of DL Quality HOs	DL Quality HOs / Total number of HOs
30	Int_HO_UL	Percentage of UL Interference HOs	UL Interference HOs / Total number of HOs
31	Int_HO_DL	Percentage of DL Interference HOs	DL Interference HOs / Total number of HOs
32	Dis_HO	Percentage of Distance HOs	Distance HOs / Total number of HOs
33	In_HO	Percentage of intracell HOs (without GPRS)	Intracell HOs / Total number of HOs
RACH			
34	RACH	Percentage of RACH capacity used	RACH load / RACH capacity
Alarms			
35	Al_A	A interface alarm	
36	Al_A-bis	A-bis interface alarm	
37	Al_LAPD	LAPD alarm	
38	Al_TRAU	Transcoder alarm	
39	Al_Com	Combiner alarm	

Number	**Code**	**Name**	**Description**
40	Al_Ant	Antenna alarm	
41	Al_BTS	BTS with no transactions alarm	
42	Al_pow	Reduced output signal alarm	
43	Al_TRAU	Transcoder related alarm	
44	Al_int	Excessive interference alarm	
45	Al_pat	Path difference alarm	
46	Al_cha	Channel failure rate alarm	
47	Al_hw	HW alarm	

Table 4.11: Relations among causes, symptoms and conditions

Causes	Subcauses	Conditions	Symptoms	
Interference	UL Interference	FH	1	DCR
		PC	2	RF_dcr
		DTX	3	RF_HO_dcr
		Cell type	11	Q1_7_UL
			14	Q0_DL-Q0_UL
			18	Q567_Lev_UL
			24	IOI
			28	Qua_HO_UL
			30	Int_HO_UL
			33	In_HO
			34	RACH
	DL Interference	FH	1	DCR
		PC	2	RF_dcr
		DTX	3	RF_HO_dcr
		Cell type	12	Q1_7_DL
			13	Q0_UL-Q0_DL
			19	Q567_Lev_DL
			25	HO_F
			29	Qua_HO_DL
			31	Int_HO_DL
			33	In_HO
Coverage	Borders	Cell type	1	DCR
		Ant.align.	2	RF_dcr
			3	RF_HO_dcr
			15	RXLEV_UL
			16	RXLEV_DL
			17	Link_imb
			21	TA_10
			26	Lev_HO_UL
			27	Lev_HO_DL
			28	Qua_HO_UL
			29	Qua_HO_DL
	Shadow zones	Cell type	1	DCR
		Ant.align.	2	RF_dcr
			3	RF_HO_dcr
			15	RXLEV_UL
			16	RXLEV_DL
			17	Link_imb
			22	TA_4
			26	Lev_HO_UL
			27	Lev_HO_DL
			28	Qua_HO_UL
			29	Qua_HO_DL
	Excessive TA	Cell type	1	DCR
		Ant.align.	15	RXLEV_UL
			23	TA_50

Causes	Subcauses	Conditions	Symptoms	
Hardware	TRXn: BB		1	DCR
			11	Q1_7_UL
			12	Q1_7_DL
			20	Q5X6X7_TRXn
			43	Al_TRAU
			46	Al_cha
	TRXn: general		1	DCR
			11	Q1_7_UL
			12	Q1_7_DL
			16	RXLEV_DL
			17	Link_imb
			41	Al_BTS
			42	Al_pow
			46	Al_cha
	Combiner		1	DCR
			2	RF_dcr
			12	Q1_7_DL
			29	Qua_HO_DL
			39	Al_Com
	Antenna		1	DCR
			2	RF_dcr
			40	Al_Ant
	TX path	Climate	1	DCR
			2	RF_dcr
			3	RF_HO_dcr
			12	Q1_7_DL
			13	Q0_UL-Q0_DL
			16	RXLEV_DL
			19	Q567_Lev_DL
			29	Qua_HO_DL
			31	Int_HO_DL
			33	In_HO
	RX path	Climate	1	DCR
			2	RF_dcr
			11	Q1_7_UL
			14	Q0_DL-Q0_UL
			15	RXLEV_UL
			17	Link_imb
			18	Q567_Lev_UL
			28	Qua_HO_UL
			30	Int_HO_UL
			33	In_HO
			45	Al_pat
	Other HW fault		1	DCR
			47	Al_hw

Causes	Subcauses	Conditions	Symptoms	
Transmission	A-bis fault		1	DCR
			4	A-bis_dcr
			5	A-bis_HO_dcr
			10	LAPD_dcr
			36	Al_Abis
			37	Al_LAPD
	A fault		1	DCR
			6	A_dcr
			7	A_HO_dcr
			35	Al_A
	Transcoder		1	DCR
			8	Tr_dcr
			9	Tr_HO_dcr
			38	Al_TRAU
Others	Fading	FH	1	DCR
		Cell type	11	Q1_7_UL
		Rec.Div.	12	Q1_7_DL
			18	Q567_Lev_UL
			19	Q567_Lev_DL
			28	Qua_HO_UL
			29	Qua_HO_DL
	Adjacencies		1	DCR
			2	RF_dcr
			3	RF_HO_dcr
			15	RXLEV_UL
			16	RXLEV_DL
			21	TA_10
			25	HO_F
	Parameters		1	DCR
			2	RF_dcr
			3	RF_HO_dcr
			25	HO_F
			26	Lev_HO_UL
			27	Lev_HO_DL
			28	Qua_HO_UL
			29	Qua_HO_DL
			30	Int_HO_UL
			31	Int_HO_DL
			33	In_HO

Clarifications:

- **Interference:**

  Symptoms: Interference may show up in the uplink path, in the downlink path or in both links. Many samples in the interfered link present high signal level but bad quality. The implication is that, although the received signal level from the serving cell is enough, interference is considerable. Therefore, $\frac{C}{I}$ ratio is reduced and, consequently, quality is degraded. Quality has been measured as the share of samples out of RXQUAL band 0.

  Conditions: If the features FH, power control or DTX are activated, it is less probable that the cause of problems is interference than if those features are not active. Furthermore, if traffic density is low (rural areas) it is less probable that the cause of problems is interference than if the traffic density is high (urban areas).

- **Lack of coverage:**

  Symptoms: Lack of coverage, either at the cell borders or at shadow areas, reduces the signal level and quality. Normally the limiting path in this type of fault is the uplink. If the lack of coverage is taking place at the cell borders, probably the cell will be large and, therefore, the average distance between the MSs and the BTS will be high. On the contrary, if the lack of coverage is due to shadow areas, probably the average distance between the MSs and the BTS will be small.

  Conditions: Lack of coverage at the borders is common in rural cells. On the other hand, shadow areas normally appear in urban areas because of the presence of obstacles, indoors, etc. If a visual inspection of the antenna shows that it is misaligned, it is very probable that the cause of the problems is lack of coverage. Similarly, if the antenna downtilt is excessive, probably the cause of the problems is lack of coverage at the borders.

- **Excessive TA**

  Symptoms: If too many calls are dropped because of the excessive distance, the percentage of MSs located at far distances from the BTS (and consequently with high values of TA) is probably high.

  Conditions: If the antenna is not downtilted enough, it is probable that the signal reaches more than 35 Km. The "Cell Type" is also related to this cause because this type of fault is more frequent in rural cells.

- **TRXn: Baseband Unit:**

  Symptoms: Depending on whether the fault is related with the reception or the transmission path, the share of samples with bad quality in the uplink or in the downlink, respectively, rises. When there is a TRX fault, the percentage of samples in RXQUAL bands 5, 6 or 7 in the uplink for the faulty TRX goes up. The quality distribution per band is different in the case of interference fault compared to TRX fault. Whereas if the problem is caused by interference the percentages of samples in each upper band increase in a similar way, when there is a TRX fault, the percentage of samples rises mainly in one single band.

- **TRXn: general fault:** This type of TRX fault is any fault in a TRX not related only with the baseband unit.

  Symptoms: The transmitted signal level decreases because of the reduction in the output signal of the faulty TRX. As a consequence, the signal level is normally higher in the uplink than in the downlink path. In addition, the quality in the downlink or uplink path worsens, due to a fault in the transmitter or the receiver part, respectively, of the TRX.

- **Transmission and reception path:**

  Conditions: These faults are more frequent in areas with hard weather, e.g. places where rain is abundant and frequent.

- **Fading:**

  Conditions: If FH is activated or reception diversity is implemented, the probability of fading diminishes. Furthermore, it is more probable that fading happens in urban cells than in rural cells because in the former case, speed is slower and obstacles which cause multipath are more numerous.

- **Configuration parameters:**

  Symptoms: Symptoms for this cause can differ a lot, depending on the type of parameter that is erroneous. The most common failures are the ones in the table above.

4.A Appendix: Case Study

In order to illustrate the concepts explained in the present chapter, in this appendix some practical troubleshooting cases from a live GSM/GPRS will be described [14]. These are specially complex cases where the ultimate cause of the problem was determined by TS experts. The diagnosis model proposed in the present chapter only includes "high level" causes. Hence, subcauses should also be considered in the model in order to use it for the cases presented hereafter. For example, according to the proposed model, Case 1 (Section 4.A.1) would be diagnosed as "UL interference". In order to diagnose the ultimate cause of the interference, i.e. the antenna back lobes were large, a more complex model would be required.

4.A.1 Case 1: UL Interference

Problem description

DCR is very high on sector 1 of a tri-sectorial site. The antennas are located on a high building in the middle of a big city. The antenna of sector 1 is higher than the others. Baseband FH is being used. The distance between macrocells in this area is less than 1 Km and all cells are handling very high traffic levels (5 or 6 TRX per cell). There are around 15 or 20 microcells or picocells below those macrocells carrying also a high traffic load.

Diagnosis actions

Uplink interference on idle time slot was very high, which in principle pinpointed to an interference fault.

A technician went to check the site and no hardware problems were found. A controller revised the alarms, but no important ones were triggered.

Troubleshooting experts checked the frequency plan in order to investigate whether the problem could be caused by interference. It was found that too many frequencies were involved. Therefore, some fine tuning trials were implemented, but the situation did not improve.

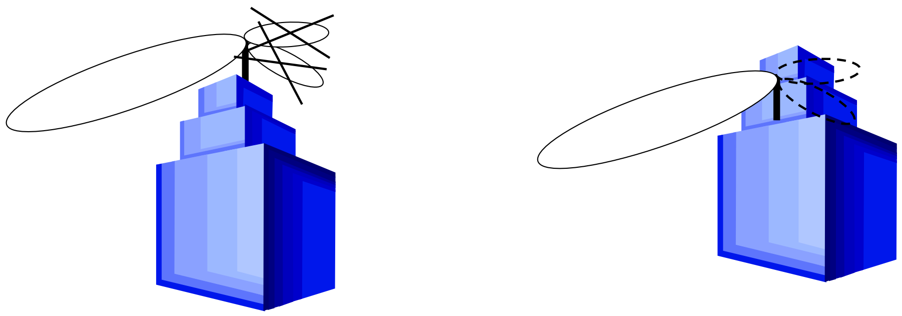
BB FH was deactivated in order to have statistics per TRX, but no TRX specific fault was found.

A technician went back to the site to measure UL interference and it was concluded that the interference was coming from MSs on other cells. Furthermore, the interference was coming from the back lobe of the antenna.

Solution

The problem was due to the fact that the antenna on sector 1 was not directional enough. Thus, the side and back antenna lobes were picking up UL signals from MSs on cells behind.

It was decided to reduce the side and back lobes of the antenna by replacing it by other antenna (Fig.4.6(a)) or moving it down on the roof (Fig.4.6(b)), so that the building could act as a screen. After implementing the first solution, the UL interference and DCR went down.



(a) Change the antenna (b) Move the antenna

Figure 4.6: Case 1: Solutions

4.A.2 Case 2: HW fault in reception path

Problem description

The DCR increases intermittently in a cell located in the middle of a medium size city. At the beginning, the problem only showed up one or two days a week. After a while, dropped calls occurred on more days of the week, and finally it has become a permanent problem. There is only an omni-directional antenna, which is used to transmit and receive. There are 6 TRXs and BB FH is used. No reception diversity is applied.

Diagnosis actions

Uplink interference on idle time slot was very high, coinciding in time of the periods with high DCR. In principle, this pinpointed to an interference fault.

A technician went to check the site and no hardware problems were found. A controller checked the alarms, but no important ones were triggered. Troubleshooting experts checked the frequency plan and they concluded that there were no problems with it.

BB FH was deactivated in order to have statistics per TRX, but no fault was found.

A technician went back to the site to measure UL interference and no unusual levels were measured. Some pieces of equipment were changed, including the antenna, but the problem remained.

The BTS power was decreased while monitoring the UL interference. It could be observed that after a certain threshold, the interference disappeared. A new duplexer was installed, but the interference was still there. The conclusion was that the interference source was located in the feeder, as the antenna was already replaced. The cable was replaced, achieving no better results. Finally, the lightning protector[18] was replaced and interference disappeared.

[18] A *lightning protector* is a device designed to divert large surges of current, such as a lightning strike, from reaching the RF equipment

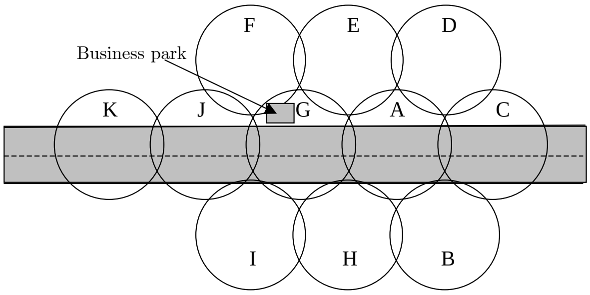


Figure 4.7: Scenario for Case 3

Solution

The problem was that there were leakages in the lightning protector to the ground when a lot of power and traffic was involved. As a result, some part of the transmitted energy was going to the ground and, after being reflected, going back to the BTS. Although the duplexer was attenuating the signal, still high power was arriving at the receiver LNA, causing intermodulation, which produced UL interference.

4.A.3 Case 3: Adjacency definition and Congestion

Problem description

The scenario for Case 3 is shown in Fig.4.7. Cell A is presenting a high DCR. Cell G is serving both a motorway and a business park. During busy hours the congestion in cell G is high.

Diagnosis actions

Cell A and D suffered from high HO Failure rates. Among the HO causes, level and quality HOs stood out .

The frequency plan was checked and it was observed that the problem existed since there had been a frequency change in an area nearby. Although D and J are cochannels, they do not show interference.

After checking the configuration parameters, it was found that the BCC[19] of cell D and J were the same one. The problem can be described as follows: A MS served by cell A driving in the direction of the business park tries to handover to cell G. However, cell G is congested. As the MS do not receive any strong signal from the other neighboring cells, the call is dropped due to lack of coverage. If case the call survives, the MS may catch the signal from cell J, which has the same frequency and BCC as cell D (cell J is not defined as adjacent of cell A). The BSC sends a HO Command expecting the MS to go to cell D. As the HO can not take place, the Failure Rate metrics rise in both A and D. After the unsuccessful HO attempt, the MS returns to cell A. Soon the lack of coverage makes the call drop and the RF_dcr counter increases. In case the MS is dropped while coming back to cell A, the counter RF_HO_dcr is increased.

[19]BCC: Base Station Colour Code. BCC is a code that identifies the base station. It is used to discriminate between cells using the same frequencies.

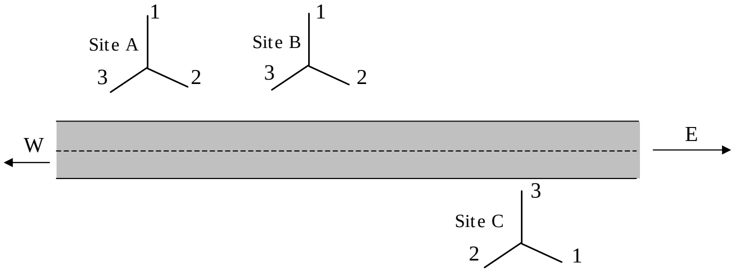


Figure 4.8: Scenario for Case 4

Solution

After the BCC of cell D or J was changed, HO_F of cell D became normal. Although HO_F on cell A also improved, due to the blocking of cell G, the value was still not good enough. Furthermore, DCR in cell A was still very high.

A new adjacency definition was added between cell A and J, which allows direct HO from A to J if cell G is congested. DCR improved, although it was still not good enough.

In order to completely solve the problem, cell G would need capacity extension (more TRXs) or a separate cell should be implemented to serve the business park.

4.A.4 Case 4: Adjacency definition, Configuration parameters and Interference

Problem description

Dropped calls are detected in a motorway (Fig.4.8). The average speed on the motorway is around 100 Km/hour.

Diagnosis actions

The problem occurred since a new site (site B), which also serves on the motorway, was added. $B1$ presented high DCR and high HO_F. However, that sector was not supposed to serve on the motorway. On cells $A2$, $B2$ and $B3$, the HO_F was also high.

From some drive tests it was found that when going east, HOs were happening from $A2$ to $B1$ around 30% of the time, instead of handing to $B2$ or $B3$. $B3$ was normally missed because MSs were moving very fast. Furthermore, due to reflections from the surrounding buildings, $B1$ was dominant in some points of the motorway. $C3$ was not defined as neighbor of $B1$, therefore although the received signal from $C3$ was high enough, the calls were dropped due to a lack of coverage.

Driving west, HOs from $C3$ to $B1$ were happening around 30% of the time. In those cases, quality diminished and finally calls were dropped. The reason was that $B1$ suffered from adjacent channel interference from cell $A3$.

Solution

$B1$ antenna was downtilted so that the cell was not dominant on the motorway. $C3$ was added as neighbor to $B1$. Finally, HO margins were also changed.

4.A.5 Case 5: DL Interference

Problem description

Cell C is showing high DCR. It is covering part of a suburban area including a road crossing.

Diagnosis actions

DL quality was worse than that of other cells in the area, whereas UL quality was similar to that of other cells in the area.

Signal level received from cell C at the road crossing was around -85 dBm. Drive tests confirmed that most drops happened at the road crossing.

Frequency plan was investigated and no issues were found.

At the building in the upper right side of the road crossing there was a site B owned by another operator. The signal received from B was around -35 dBm. The 50 dB difference between the serving level and the interfering level was too much and, therefore, MSs in cell C were suffering from interference.

Solution

A higher signal level was provided at the road crossing so that the signal difference got lower than the current 50 dB. However, the best solution would be adding a new site close to the road crossing.

4.A.6 Summary

In this appendix, some typical faults encountered in routine operation of real networks have been presented. The case studies have shown that TS normally consists of "try and error" procedures, e.g. something is changed, such as a piece of equipment or a configuration parameter, and then it is checked whether that action solved the problem. Very often, it is not clear how to identify the fault cause, which makes it impossible to deploy TS as a series of checking routines. Hence, an automatic tool which takes into account the uncertainty inherent to TS becomes essential to reduce operational expenditure.

Chapter 5

Bayesian modelling of fault diagnosis

This chapter proposes different diagnosis systems, which answer two important questions: i) how to represent the knowledge base explained in Chapter 4; ii) how to assess which is the cause of the problems, based on the known values of some symptoms and conditions.

Section 5.1 presents some concepts used in this chapter and the main elements of all the systems described in the following sections. Section 5.2 proposes a diagnosis system based on a Bayesian classifier. The systems presented in sections 5.3-5.6 are based on Bayesian Networks. The main problems faced when building a model based on BNs are summarized in Section 5.3 and they are further explained in subsequent sections. In Section 5.4, the structure of the BN is described. In Section 5.5, some algorithms to learn the parameters of the model from training data are proposed. In Section 5.6, two methods to diminish the diagnosis errors when the model parameters are inaccurate are explained. Finally, Section 5.7 explains how to build the diagnosis systems based on the knowledge from experts in troubleshooting, i.e. how to convert the knowledge base presented in Chapter 4 into the diagnosis systems proposed in the present chapter.

5.1 Introduction

5.1.1 Definitions

Two components of diagnosis systems have been distinguished: the model and the inference method. The *model* is a representation of how the "world works" in the area under study. Ideally, the model should summarize the behavior of the RAN of cellular networks. However, when the model is built by human experts, it represents the *knowledge* about the application domain, i.e. the knowledge of troubleshooting experts about how the identification of faults should be done (Fig.5.1). Hence, any model is inevitably a simplification of what the expert knows, which is itself a simplification of reality. Logically, it is not feasible to consider all possible situations that may occur and there is always a trade-off between accuracy and manageability of the model.

There are two aspects involved in the model construction: the information about the appli-

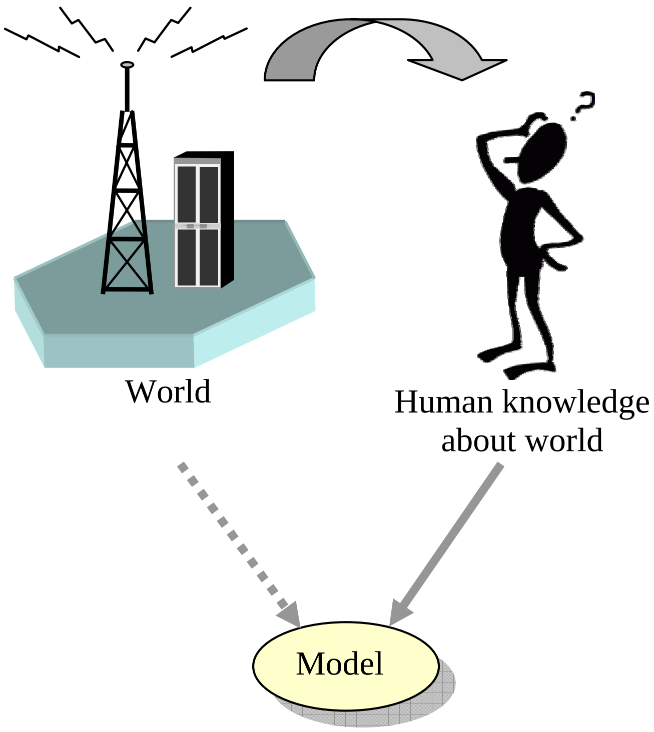


Figure 5.1: Diagnosis concepts

cation domain contained in the model and the model representation. The first item, also called *knowledge base*, was presented in Chapter 4. In the present chapter the *model representation* will be described for different diagnosis systems.

The *inference method* is the algorithm that identifies the cause of the problems based on the available evidence (e.g. symptoms). For example, an inference method would be the reasoning followed by a doctor to say, after observing that a patient has fever and he coughes, that he has flu. All inference methods presented in this thesis are based on the Bayes' rule.

According to the knowledge base explained in Chapter 4, all the diagnosis models should be composed of the following elements: causes, symptoms and conditions. The *causes* represent the possible faults that may be producing problems in the network. The *symptoms* represent manifestations of the causes. Finally, the *conditions* represent factors that have to be taken into account in order to properly identify the cause. These elements have been modelled as random variables. The inherent continuous nature of performance indicators has been one of the main difficulties found in the development of this thesis. One of the differences between the models in the following sections is whether performance indicators have been modelled as continuous or discrete variables. In the later case, a discretization method has been required.

Models should also define the relationships among the variables. In the bayesian approach adopted in this thesis, those relations are specified by means of joint probability density functions.

In summary, definition of models should include the following information:

Table 5.1: Conditions

Conditions	**States**	**Random variable**
Frequency Hopping	off, on	0, 1
Power Control	off, on	0, 1
Discontinuous Transmission	off, on	0, 1
Reception Diversity	off, on	0, 1
Cell Type	low density, medium density, high density	-1, 0, 1
Climate	wet, normal, dry	-1, 0, 1
Antenna Alignment	not ok, ok	0, 1

- Elements (causes, symptoms and conditions), that is the knowledge base.
- Random variables that model the previous elements. In the case of discrete random variables which model continuous elements, the discretization method should also be defined.
- Joint probability density functions, which stand for the relationship among random variables

The first point is the same one for all models and it was covered in Chapter 4. Furthermore, the random variables used for causes, conditions and alarms are the same for all models. However, the type of random variables used to model performance indicators and the assumptions adopted to define the joint pdfs are different for each model presented in the following sections.

In order to simplify, **causes** have been modelled as binary discrete random variables. Therefore, it will be assumed that a cause is a random variable with two possible values: $0(no)$, $1(yes)$, related to the absence or presence of the cause, respectively. Causes considered in the model are the ones in Table 4.1. A more realistic model would consider causes as continuous variables, depending on the degree of seriousness of the fault. An intermediate solution would be to discretize causes into a limited number of states, e.g. very serious, serious, acceptable, slight, non-existent.

Conditions have also been represented as discrete random variables because conditions included in the model have a limited number of states. For example, the condition representing FH can take the values 1 or 0, depending on whether this functionality is activated in the network or not. Conditions considered in the model are shown in Table 5.1. The table also shows the states of the conditions and the possible values of the corresponding random variables.

Alarms have been modelled as binary random variables, whose possible values are 0 or 1, corresponding to the states *off* or *on*, respectively. Alarms were enumerated in Table 4.6.

Performance indicators were shown in Table 4.10. They can adopt continuous values in the range $[0-100]$% (except symptoms 13 and 14, which can go from -200 to 200 % and symptom 17, which is in the range $[0-200]$%). Performance indicators are continuous random variables in the model presented in Section 5.2, whereas they are discretized in the models described in sections 5.3-5.6.

5.1.2 Notation

The number of states of random variable X_i is denoted by $|X_i|$. The states of a variable X_i are $\{x_{i,1}, ..., x_{i,|X_i|}\}$. The symbol x_i is used to represent any value of X_i, $x_i = x_{i,j}$, $j \in \{1, ..., |X_i|\}$. Evidence $E = \{X_1 = x_1, ..., X_P = x_P\}$ is an assignment of values to P variables. The value of a variable X_i in assignment E is denoted by x_i^E. Variables can be causes, symptoms or conditions.

$C = \{C_1, ..., C_K\}$ is the set of causes. When all causes are modelled as a single random variable, the variable is called C and its states, $\{c_1, ..., c_K\}$, are the different causes. When different random variables are distinguished for each cause, they are called $C_1, ..., C_K$ and their states are *no*/*yes*. In both cases, in order to refer to any of the causes, the notation c will be used.

$S = \{S_1, ..., S_M\}$ is the set of symptoms. The states of symptom S_i are $\{s_{i,1}, ..., s_{i,Q_i}\}$. $O = \{O_1, ..., O_L\}$ is the set of conditions. The states of condition O_i are $\{o_{i,1}, ..., o_{i,H_i}\}$.

X_i^C is a continuous symptom which has been discretized into the discrete symptom X_i. The interval limits (thresholds) of X_i^C used in the discretization are $\{t_{i,1}, ..., t_{i,T_i}\}$. Probability of a variable X_i taking a value $x_{i,j}$ is denoted by $P(X_i = x_{i,j}) = P(x_{i,j})$.

A *case* is a set of values for some symptoms and conditions. A case can also include the real cause, c. D is a set of N cases, with the m^{th} case given by the $(P+1)$-tuple $x^{(m)} = \{c^{(m)}, x_1^{(m)}, ..., x_R^{(m)}, o_1^{(m)}, ..., o_{P-R}^{(m)}\}$.

The complementary set of $A \subset U$ is $B = U \backslash A$.

5.2 Model based on Bayesian Classifier

Classification, which is a basic task in data analysis, is the action of assigning an unlabelled example (described by a set of attributes) to a class. A *classifier* is the algorithm that carries out the classification. The class label is the name identifying the class.

Diagnosis can be considered as a classification problem typical in machine learning applications [140], where the classes would be the causes and the attributes would be the symptoms. One of the most effective classifiers, in the sense that its performance is competitive with state-of-the-art classifiers, while being very simple, is the **Naive Bayesian Classifier (BC)**. It has proven to be effective in many practical applications, including text classification, medical diagnosis, and system performance management [165].

Because of the previous reasons, the fist diagnosis system for the RAN of cellular networks proposed in this thesis is based on a BC [46]. In Section 5.2.1 the inference method for the BC will be described. In Section 5.2.2 the known beta pdf will be revised. Then, the model, which is based on beta pdfs, will be presented in Section 5.2.3.

5.2.1 Inference Method

The purpose of a classifier is to predict the value of a discrete class variable C, based on the known values of a set of attributes $S = \{S_1, ..., S_M\}$. The possible values of the class, $C = \{c_1, ..., c_K\}$, are the different causes. The attributes are the symptoms, which are modelled as continuous random variables. Let's consider a *case* $E^{(m)} = \{S_1 = s_1^{(m)}, ..., S_M = s_M^{(m)}\}$, composed of the values for each of the M symptoms (or a subset of the M symptoms).

In order to statistically define the model completely, the joint pdf of the random variables in the model would be required. This task becomes unaffordable if the number of variables is large and the set of training cases is scarce. The BC simplifies the design of the model by means of the following assumptions:

1. Only one cause is present at a time (due to the fact that a single variable is used to model all causes)
2. Symptoms are independent of conditions given the cause.
3. Symptoms are independent given the cause.

In mobile networks, the first two assumptions are reasonable because they are true in most cases. The third assumption, the independence of symptoms given the cause, is often unrealistic in GSM networks. Nevertheless, it has been demonstrated that even if strong attribute dependencies exist the BC performs very well [165, 74, 82]. Furthermore, the BC has many advantages, such as its simplicity, learning and classification speed.

A classifier takes a case $E^{(m)}$, it computes a set of discriminant functions of the case, $f_i^{(m)}$, $i = 1, ..., K$, one for each class, and it assigns the case to the class whose function is maximum. For the BC the discriminant functions are

$$f_i^{(m)}\left(E^{(m)}\right) = P\left(c_i\right) \prod_{j=1}^{M} P\left(S_j = s_j^{(m)} \middle| c_i\right) \quad , \quad i = 1, ..., K \tag{5.1}$$

where $P(c_i) = P(C = c_i)$ is the prior probability of the class c_i, and $P(S_j = s_j^{(m)}|c_i)$ is the conditional probability of symptom $S_j = s_j^{(m)}$ given class c_i. In other words, the BC finds the *maximum a posteriori* probability (MAP) class given the attributes.

This classifier can be derived from Bayes' theorem if it is assumed that the attributes are independent given the class. In that case

$$P\left(c_i \middle| E^{(m)}\right) = \frac{P\left(c_i\right) \prod_{j=1}^{M} P\left(S_j = s_j^{(m)} \middle| c_i\right)}{P(E^{(m)})} \tag{5.2}$$

where $P(E^{(m)})$ is the probability of the case and it can be ignored to calculate the maximum because it is the same for all classes.

Consequently, automated diagnosis can be performed by calculating the probability of each cause given a set of symptoms and conditions as follows

$$P(c_i|E^{(m)}, O^{(m)}) = \frac{P(c_i|O^{(m)}) \prod_{j=1}^{M} f_{S|C}(S_j = s_j^{(m)}|c_i)}{f_{S|O}(E^{(m)}|O^{(m)})} =$$

$$= a \cdot P(c_i|O^{(m)}) \prod_{j=1}^{M} f_{S|C}(S_j = s_j^{(m)}|c_i) \tag{5.3}$$

where $E^{(m)}$ is a vector of evidence for M symptoms, $O^{(m)}$ is a vector of evidence for L conditions, a is a constant and some conditional probabilities have been substituted by probability density functions because most symptoms are continuous.

The cause with the highest posterior probability is assumed to be the one that is causing the high number of dropped calls in a cell. Therefore, the inference method for this diagnosis system consists of calculating the fault cause as

$$\max_{c_i}[P(c_i|E^{(m)}, O^{(m)})] \tag{5.4}$$

In order to apply (5.4) the probability of causes given conditions and the pdfs of symptoms given causes are required. Therefore, defining the model for diagnosis means specifying the qualitative part (causes, symptoms and conditions) and the quantitative part (probability functions). On the one hand, the qualitative part can be easily provided by experts. On the other hand, the probabilities can be estimated from a database of labelled examples. However, in most cases the amount of data available is very limited and calculating the conditional pdfs is not possible, especially if nothing is previously known about them. This has been tackled in the bibliography by discretizing the continuous attributes or assuming that the pdfs are gaussian [75, 80]. In Section 5.2.3 the model for diagnosis will be presented and it will be shown that, due to the characteristics of the symptoms, the definition of the conditional pdfs consists of setting some parameters for a known distribution, which will be introduced in the following section.

5.2.2 Beta distribution

Firstly, attention should be drawn to two known distributions: Bernoulli and beta distributions:

- Let's suppose an experiment whose outcome can be either success (1) or failure (0). It is said that a random variable X, related to the previous experiment, follows a *Bernoulli distribution* with parameter β ($0 \leq \beta \leq 1$) if X can take only the values 0 and 1 with probabilities $1 - \beta$ and β, respectively.

- A *beta random variable*[1] with parameters (a, b) has the following probability density func-

[1] A Bernoulli experiment, with probability of success β, is repeated n times, obtaining k successes. Let's suppose that β is a random variable and $f(\beta)$ its probability density function. Let θ be the event "obtaining k successes". The probability of θ is $P(\theta|\beta) = \beta^k(1-\beta)^{n-k}$. Applying Bayes'theorem:

$$f(\beta|\theta) = \frac{\beta^k(1-\beta)^{n-k} f(\beta)}{\int_0^1 \beta^k(1-\beta)^{n-k} f(\beta) d\beta} \tag{5.5}$$

Assuming that β is uniform in the interval (0,1), the previous equation becomes [155]:

$$f(\beta|\theta) = \frac{(n+1)!}{k!(n-k)!} p^k(1-p)^{n-k} \tag{5.6}$$

If this equation is generalized and n and k can be not only integers, but also real numbers, then this function is known as *beta density function*.

tion:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}\ ,\ 0 \le x \le 1 \tag{5.7}$$

where a and b are positive real numbers and $\Gamma(x)$ is the gamma function:

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \tag{5.8}$$

The mean and variance of the beta function are

$$\mu = \frac{a}{a+b} \qquad \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)} \tag{5.9}$$

The uniform distribution is a special case of the beta distribution when $a = b = 1$.

The Bernoulli and beta distributions are related as will be explained hereafter. Let $x_1, ..., x_n$ be a random sample from a Bernoulli distribution for which the value of the parameter β is unknown ($0 < \beta < 1$). Suppose that the prior distribution of β is a beta distribution with parameters a and b. Then the posterior distribution of β, given that the random sample $x_1, ..., x_n$ is observed, is a beta distribution with parameters $a + y$ and $b + n - y$, $y = \sum_{i=1}^{n} x_i$ [70, 156]. That is, if the prior distribution of β is a beta distribution, then the posterior distribution at each stage of sampling will also be a beta distribution, regardless of the observed values in the sample.

An application of the previous theorem is given in the following example [70]. Suppose that the proportion θ of defective items in a large shipment is unknown; that the prior distribution of θ is a beta distribution with parameters a and b; and that n items are selected one at a time at random from the shipment and inspected. If the first inspected item is defective, the posterior distribution of θ will be a beta distribution with parameters $a + 1$ and b. If the first item is not defective, the posterior distribution will be a beta distribution with parameters a and $b + 1$.

Beta pdfs have a long tradition representing prior beliefs concerning a relative frequency, e.g. in medical diagnosis [156]. Different arguments to support the use of the beta pdfs for priors can be found in the bibliography. First, if a uniform pdf (or more generally, a beta pdf) is used to represent prior beliefs, after some data is known the posterior density function is also beta. The second argument was given in [203]: if certain assumptions about an individual's belief are made, then that individual must use the beta density function to quantify any prior belief about a relative frequency. The demonstration was done for the Dirichlet distribution, which is a generalization of the beta distribution to the case of more than two alternatives.

If we assume that our prior belief follows a beta pdf, there are different ways of estimating the values of a and b. According to (5.9), the ratio between a and b determines the mean, e.g. when a is large compared to b, the mean will be close to 1. On the other hand, for fixed mean the sum of a and b determines the spread of the distribution, that is the variance. For example, Fig.5.2 shows beta pdfs for different values of the parameters a and b.

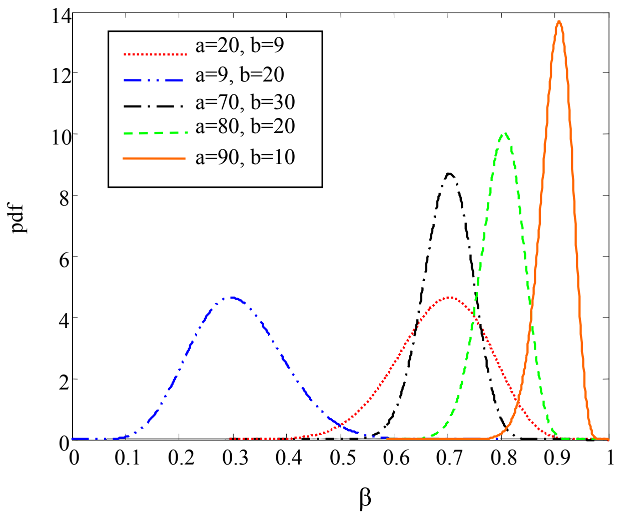


Figure 5.2: Beta probability density functions

5.2.3 Model Representation

In the diagnosis system based on the BC, performance indicators have been modelled as continuous random variables. Furthermore, as explained in 5.2.1, the quantitative part of the model is composed of the following probabilities:

1. Probability of each cause given that there is a problem (high DCR) and given the conditions. The probabilities of the causes are not required for all the possible combinations of conditions because normally only some of them are possible.

2. Probability density functions of the symptoms given each cause.

The first probabilities can easily be elicited by experts in troubleshooting. In the second group of probabilities the two types of symptoms (performance indicators and alarms) should be distinguished. Firstly, the pdf for an alarm given the causes is reduced to the probability of the alarm being active given each cause, which, if the number of considered alarms in the model is low, can also be obtained from the experts' knowledge. Nevertheless, defining the conditional pdfs for the performance indicators is much more complex because these symptoms are modelled as continuous variables. Therefore, either experts should specify the pdfs or performance indicators should be discretized so that probabilities are more easily elicited by experts.

However, there is an alternative solution: if the symptoms followed a known pdf, defining the pdf would become a parameter estimation problem. In order to find an appropriate known pdf for the symptoms, one could consider how the symptoms related to performance indicators were defined: most of them can be obtained from the percentage of samples complying with a given condition C (the special cases that do not follow this assumption will be studied below). In other words, they are the relative frequency of one of the two possible outcomes of an experiment (condition C achieved or not), which can be described by a Bernoulli random variable X. For example, for the symptom "Percentage of samples in UL with RXLEV<10", the condition is "RXLEV<10", therefore a random variable X with two values, 0 and 1, corresponding to RXLEV<10 and RXLEV≥10, respectively, can be defined. Then, let Y be another random variable, having real values in the interval [0,1], representing the expert's belief concerning the relative frequency with which $X = 1$ if a cell and day is randomly chosen. The pdf of Y can be seen as the prior belief in the β parameter of the Bernoulli distribution followed by X. In Section 5.2.2, it was explained that under these conditions the most adequate pdf is the beta pdf. The beta pdf has previously been used to represent beliefs concerning a relative frequency [143] in medical diagnosis [156], land cover proportions in geological models [61], frequency distributions of solar radiation indexes [89], etc.

The conclusion is that the pdfs of most of the symptoms given the causes in the diagnosis model can be approximated by beta pdfs [46]. Therefore, the only required information to completely define the conditional probabilities in the model are the parameters a and b of the beta pdfs for each pair symptom/cause. Those parameters can either be elicited by experts [143, 148] or they can be obtained from data by means of appropriate techniques [164], e.g. maximum likelihood estimation.

However, some symptoms in the model are not percentages of samples. Their pdfs will be described in the following paragraphs.

Symptoms 13 and 14

Symptoms 13 and 14 in Fig.5.3, DL and UL quality link imbalance, are the difference between two random variables which can be modelled as beta distributions (symptoms 11 and 12).

If X and Y are two independent random variables with pdfs f_X and f_Y, respectively, then the pdf f_Z of $Z = X + Y$ is given by the convolution of the pdfs f_X and f_Y [70]:

$$f_Z(x) = \int_{-\infty}^{\infty} f_X(x-y) f_Y(y) dy \tag{5.10}$$

On the other hand, if $Q = -Y$, then

$$f_Q(x) = f_Y(-x) \tag{5.11}$$

The pdf $f_{S_{13}}(x)$ of $S_{13} = S_{12} - S_{11}$ can be calculated based on (5.10) and (5.11) as

$$f_{S_{13}}(x) = \int_{-\infty}^{\infty} f_{S_{12}}(x+y) f_{S_{11}}(y) dy \tag{5.12}$$

which is the correlation of $f_{S_{12}}$ and $f_{S_{11}}$.

Taking into account that the random variables for symptoms 13 and 14 keep the relation $S_{13} = -S_{14}$, the pdf for symptom 14 is:

$$f_{S_{14}}(x) = f_{S_{13}}(-x) = \int_{-\infty}^{\infty} f_{S_{11}}(x+y) f_{S_{12}}(y) dy \tag{5.13}$$

Fig.5.3 shows the pdfs of symptoms 11 and 12, modelled as beta pdfs, and the corresponding

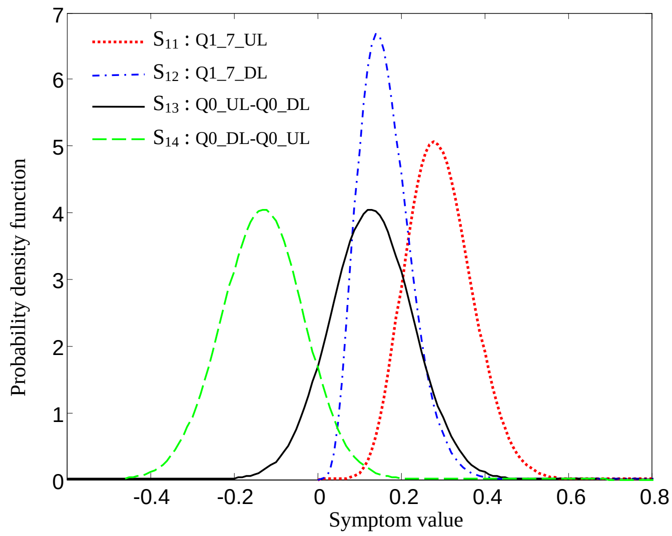


Figure 5.3: Probability density functions for difference of betas

pdfs of symptoms 13 and 14, obtained as their correlation.

Symptom 17

Symptom 17, link imbalance, is the absolute value of the difference of two random variables, $S_{17} = abs(S_{16} - S_{15})$. In the previous section, the pdf of the difference of two beta random variables has been calculated. On the other hand, it can be easily demonstrated that the pdf of $Y = |X|$ is:

$$f_Y(x) = f_X(x) + f_X(-x) \;\; ; \;\; 0 \leq x \leq \infty \tag{5.14}$$

Therefore, the pdf for symptom 17 is:

$$f_{S_{17}}(x) = \int_{-\infty}^{\infty} f_{S_{16}}(y+x) f_{S_{15}}(y) dy + \int_{-\infty}^{\infty} f_{S_{15}}(y+x) f_{S_{16}}(y) dy \;\; ; \;\; 0 \leq x \leq \infty \tag{5.15}$$

Fig.5.4 shows the pdfs of symptoms 15 and 16 modelled as beta pdfs and the resulting pdf for symptom 17.

Symptom 20.n

Symptom 20 is calculated for each of the n TRXs of the cell, as the maximum of three values: percentages of samples with RXQUAL in band 5, 6 or 7 in TRX number n. It can be demonstrated that if X, Y and Z are independent random variables with pdfs f_X, f_Y and f_Z, the pdf of $W = max(X, Y, Z)$ is:

$$f_W(x) = f_X(x)F_Y(x)F_Z(x) + F_X(x)f_Y(x)F_Z(x) + F_X(x)F_Y(x)f_Z(x) \tag{5.16}$$

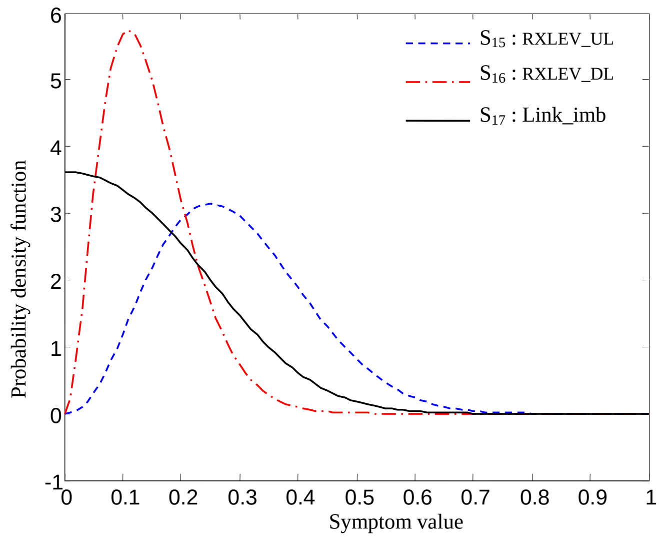


Figure 5.4: Probability density functions for symptoms related to *Link imbalance*

where F_X, F_Y and F_Z are the distribution functions of X, Y and Z, respectively.

Therefore, the pdf for symptom 20.n can be calculated as

$$f_{S_{20.n}} = f_{B5}(x)F_{B6}(x)F_{B7}(x) + F_{B5}(x)f_{B6}(x)F_{B7}(x) + F_{B5}(x)F_{B6}(x)f_{B7}(x) \tag{5.17}$$

where f_{B5}, f_{B6} and f_{B7} are the beta pdfs and F_{B5}, F_{B6} and F_{B7} their corresponding distribution functions of the three symptoms defined as follows:

$B5$=Percentage of DL samples with RXQUAL in band 5 in TRX n

$B6$=Percentage of DL samples with RXQUAL in band 6 in TRX n

$B7$=Percentage of DL samples with RXQUAL in band 7 in TRX n

Fig.5.5 shows the pdfs of $B5$, $B6$ and $B7$ modelled as beta pdfs and the resulting pdf for symptom 20.n.

5.3 Models based on Bayesian Networks

Bayesian Networks (BNs) [110, 157], also called Probabilistic Belief Networks, were described in Chapter 3. The study of inference methods for BNs is out of the scope of this thesis. Inference methods have been widely studied in existing bibliography [110, 54, 69, 119]. Thus, efficient inference algorithms exist to obtain the probability of a certain variable given the available evidence, and they are integrated in comercial tools for BNs.

The model representation is composed of variables (causes, symptoms and conditions) and relations among them. The relationships are represented as directed edges among variables and probability density functions. The main problems encountered during model construction have been the definition of the BN structure, the modelling of continuos symptoms and the specification of the probabilities in the BN:

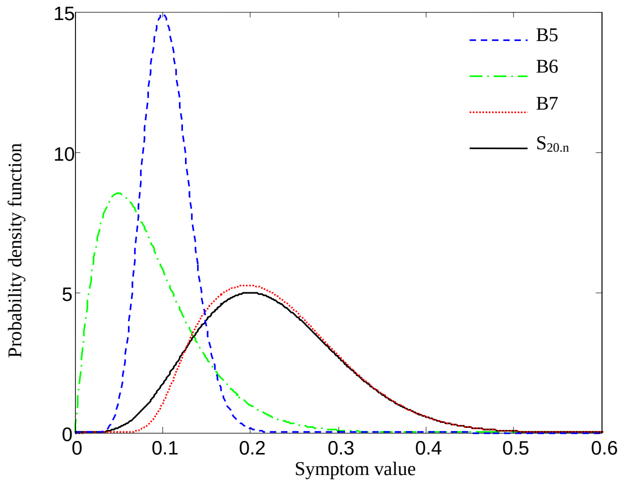


Figure 5.5: Probability density functions for maximum of betas

- One of the main drawbacks of BNs is the difficulty in the definition of complex models. Thus, a technique often used in order to simplify model construction is to assume a given network structure. That is, a certain independence between the variables is assumed. In that way, the problem of BN definition is simplified to the specification of the nodes and pdfs of a given network. The benefits of a simple network structure are not only that knowledge acquisition is easier, but also that inference algorithms are simpler. The network structures that have been selected in this thesis have taken into account that the simplicity of the diagnosis system is a key issue in the diagnosis of cellular networks [45]. The following section will describe the different structures adopted in this thesis for the development of diagnosis systems for the RAN.

- For most applications, discrete BNs are normally preferred to BNs with continuous variables. The reason is the high complexity of the knowledge acquisition and inference algorithms in the later case. Hence, the diagnosis systems proposed hereafter are based on discrete BNs. Therefore, the continuous performance indicators should be discretized before a discrete BN can be used. In Section 5.5.1 some discretization policies, either defined by troubleshooting experts or obtained based on training classified examples, will be presented. Furthermore, some novel methods which combine human expertise and information from previous cases will also be proposed.

- Once continuous variables have been discretized, the definition of the continuous pdfs in the BN is converted into the specification of probability tables. Those probabilities can be either elicited by troubleshooting experts or obtained based on training cases. Different methods for probability elicitation will be proposed in Section 5.5.2.

- Contrary to other application domains, such as medical diagnosis, for which repositories of machine learning databases have been created and maintained [50], in cellular systems there are no databases of classified cases. This is the reason why in most cases the model has to be based on knowledge, i.e. experts elicit the model parameters (probabilities in the BN and thresholds for discretized variables). When the number of parameters to be defined is large, which is normally the case, inaccuracy in the parameters is unavoidable due to multiple reasons. Firstly, experts in troubleshooting are not used to the terminology used in BN theory. Secondly, even when experts are used to build BNs, eliciting the model parameters is an arduous and very time-consuming task. Furthermore, cells operating in different environments (rural/urban, indoor/outdoor, top of building/street corner, etc.) display different characteristics, thus requiring different parameters in the model. Lastly, even similar types of cells may present different behaviors, depending on the different configuration parameters in the network (the number of configuration parameters is so large that it is not viable to model of all them in the BN). In this thesis, two methods to diminish diagnosis error of models due to inaccurate parameters are proposed. Those methods, which have been named Smooth Bayesian Networks and Multiple Uniform Intervals, will be described in Section 5.6.

5.4 Bayesian Network Structures

In this section, three BN structures will be proposed for the diagnosis models. It should be noticed that mobile network operators consider that a feasible diagnosis system should be easy to design, operate and maintain. Thus, there is a trade-off between accuracy and simplicity of the BN structure. The first model is the Simple Bayes Model (SBM), which has been chosen because of its simplicity and good performance. The following structure, the Central Bayes Model (CBM), is a novel modification of the SBM. Finally, the last proposed structure, independence of causal influence (ICI) models, is based on causal independence assumptions.

5.4.1 Simple Bayes Model (SBM)

One of the first applications of the Simple Bayes Model, also called Naive Bayes Model, was the diagnosis of congenital heart disease [191]. Since then, the SBM has been successfully utilized in many applications because of its simplicity and accurate results.

Model representation

The SBM consists of a single parent node C and M children nodes $S_1, ..., S_M$ (Fig.5.6). The states of the parent node are the possible causes $C = \{c_1, ..., c_K\}$, whereas the children are the symptoms, which may take any number of states. In the BNs developed in this thesis, in order to simplify knowledge acquisition, symptoms have two or, as much, three states.

Associated to each symptom S_i there is a table of conditional probabilities $P(S_i|C)$ of size $|C| \times |S_i|$. Likewise, associated to the cause C there is a table of prior probabilities $P(C)$ of size $|C|$.

It should be pointed out that the dependencies coded by the SBM coincides with those of the BC presented in Section 5.2. However, due to the fact that only discrete BNs have been

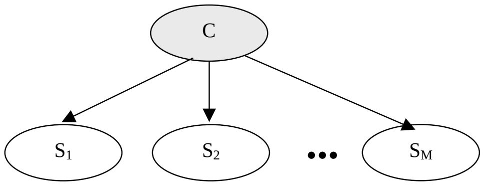


Figure 5.6: Simple Bayes Model

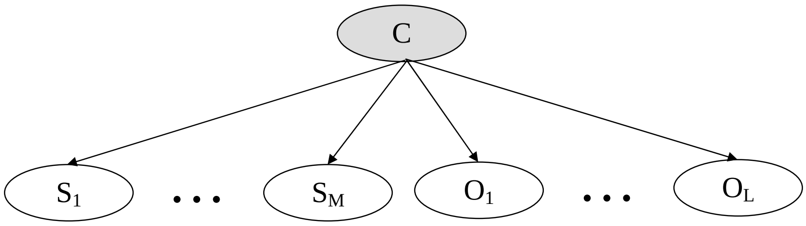


Figure 5.7: Simple Bayes Model including conditions

considered, when talking about the SBM it will be assumed that we will be referring to a discrete SBM. In the model based on the BC the main difficulty was defining the pdfs of symptoms, which were continuous, given the causes. However, in the model based on the discrete SBM, discretization of continuous symptoms and definition of probabilities are the most complex tasks.

As occurred with the BC, the following assumptions are inherent to the SBM:

- Single fault assumption, that is only a fault cause can be present at the same time. This is due to the fact that causes are the mutually exclusive states of the node C
- Symptoms are independent given the cause. It implies that once the cause is known, information about symptom S_i is of no relevance to predict the probability of other symptom S_j, $i \neq j$

The first assumption is reasonable because in the RAN normally one single cause is present at a time. As stated in 5.2.3, although the second assumption is not justified in the RAN, some studies have shown that even if strong dependencies exist among the symptoms, the SBM provides good results [165, 74, 82].

Conditions $O_1, ..., O_L$ can be modelled as children of the parent node as suggested in [173] (Fig.5.7). This solution does not respect the causal relations among conditions and causes, which in principle, should not be a problem. This model assumes that conditions are independent given the causes. The problem with this representation is that knowledge acquisition is more complex, as it will be shown in 5.7.1.

Inference method

According to the SBM, the probability of a cause c_i can be easily calculated by applying the Bayes' rule. If some evidence is available, i.e. the value of some symptoms $E^{(m)} = \{S_1 =$

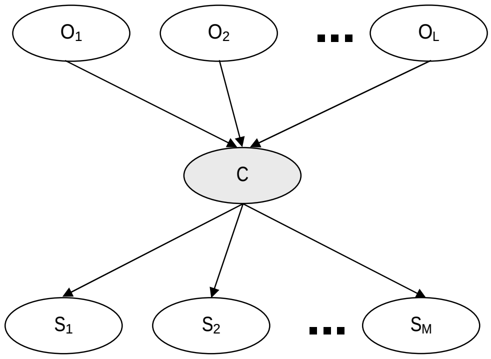


Figure 5.8: Central Bayes Model

$s_1^{(m)}, ...S_M = s_M^{(m)}\}$ and conditions $O^{(m)} = \{O_1 = o_1^{(m)}, ...O_L = o_L^{(m)}\}$ are known, the probability of the cause can be obtained as:

$$P(c_i|E^{(m)}, O^{(m)}) = \frac{P(c_i) \cdot \prod_{j=1}^{M} P(S_j = s_j^{(m)}|c_i) \cdot \prod_{j=1}^{L} P(O_j = o_j^{(m)}|c_i)}{P(E^{(m)}, O^{(m)})} \tag{5.18}$$

5.4.2 Central Bayes Model (CBM)

Model representation

Modelling conditions as children of the parent cause, as proposed in the previous model, has some problems. Firstly, the assumption that conditions are independent given the cause is often incorrect. Secondly, transformation of probabilities provided by troubleshooting experts into the required probability tables is not straightforward.

In order to overcome those problems, a structure, which has been named Central Bayes Model (Fig.5.8) is proposed. The CBM is composed of a cause node C, whose states are the possible causes, children nodes representing the symptoms $S_1, ..., S_M$ and parent nodes of C modelling the conditions $O_1, ..., O_L$.

Associated to the cause node C there is a probability table of size $|C| \prod_{j=1}^{L} |O_j|$. If the number of conditions in the model or the number of states of condition nodes are large, the previous probability table becomes intractable. However, normally the number of conditions included in the model for diagnosis in the RAN is small and so is the resulting probability table.

The assumptions inherent to the SBM (single fault and symptoms independency given cause) are still valid. Furthermore, according to the model, symptoms are independent of conditions given the cause.

Inference method

According to the CBM, if some evidence are available, i.e. the value of some symptoms $E^{(m)} = \{S_1 = s_1^{(m)}, ...S_M = s_M^{(m)}\}$ and conditions $O^{(m)} = \{O_1 = o_1^{(m)}, ...O_L = o_L^{(m)}\}$ are known, the probability of the cause c_i can be obtained as:

$$P(c_i|E^{(m)}, O^{(m)}) = \frac{P(c_i|O^{(m)}) \cdot \prod_{j=1}^{M} P(S_j = s_j^{(m)}|c_i)}{P(E^{(m)}|O^{(m)})} \tag{5.19}$$

which is the discrete version of eq.(5.3).

5.4.3 Independence of causal influence (ICI)

In order to overcome the single fault assumption inherent to the SBM and CBM, each cause should be represented as an independent node with two states (*no/yes*). Fig.5.9 shows part of a BN where multiple causes, $C_1, ..., C_K$, contribute to a common effect S_j. In this model, if K is large, the conditional probability table for symptom S_j may become intractable because it should include the probability of the symptom for each combination of the parents' states. For example, if there are 10 binary causes and a binary symptom S_j, 2^{10} probabilities should be assessed. A more tractable model would only require the probabilities of the symptom given each individual cause. Then, if more than a cause is present at the same time, the probability of the symptom could be computed using some "universal" combination rule. For example, a doctor is expected to have estimates of the chances that an individual with a given disease has a high fever, but when he is asked to estimate the probability of high fever given some rare combination of diseases, the doctor will probable have problems to provide a probability.

In this situation, independence of causal influence[2] (ICI) assumptions [97, 180, 95] simplifies knowledge elicitation and inference. Causes are said to influence their common effect independently if individual contributions from different causes are independent and the total influence is a combination of the individual contributions. The number of probabilities to be defined for S_j in Fig.5.9 is linear in K when assuming ICI, whereas in an unrestricted model the number of probabilities is exponential in K.

The dependencies in an ICI model can be represented as shown in Fig.5.10. This BN structure illustrates that the causes $C_1, C_2, ..., C_K$ *independently* contribute to the effect S_j through intermediate variables $A_1, A_2, ..., A_K$ and a deterministic function g. In other words, $C_1, ..., C_K$ are said to influence S_j independently if there exist random variables $A_1, A_2, ..., A_K$ such that [207, 208]:

1. For all i, A_i depends on C_i and it is conditionally independent of all other C_j given C_i
2. There exists a commutative and associative function g over the domain of S_j such that $S_j = g(A_1, A_2, ..., A_K)$

Under these hypothesis, the conditional probability of the symptom S_j can be formalized as follows:

$$P(S_j = s_{j,k}|C_1, ..., C_K) = \sum_{g(A_1,...,A_K)=s_{j,k}} P(A_1, ..., A_K|C_1, ..., C_K) \tag{5.20}$$

[2]In the literature, independence of causal influence was previously denoted as *causal independence*

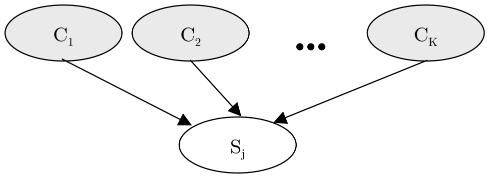


Figure 5.9: Multiple causes affecting a common symptom

Table 5.2: $g(x, y)$ for ICI example

	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	3	3
3	0	3	3	3

In addition, taking into account that each variable A_i is only dependent on its associated cause C_i, eq.(5.20) can be easily simplified as:

$$P(S_j = s_{j,k}|C_1, ..., C_K) = \sum_{g(A_1,...,A_K)=s_{j,k}} \prod_{i=1}^{K} P(A_i|C_i) \tag{5.21}$$

Therefore, in order to obtain the conditional probability of symptom S_j the only required data are the probabilities of the intermediate variables $P(A_i|C_i)$, $i = 1, ..., K$. Thus, the number of probabilities to be specified grows linearly with K.

Example 1 (adapted from [207]). Faculty members at a university are evaluated in teaching, research and management for the purpose of obtaining a salary complement. A faculty member's salary is decreased, maintained, increased or double increased depending on whether his performance is evaluated unacceptable in at least one of the three areas, acceptable in all areas, excellent in one area, or excellent in at least two areas, respectively.

Let C_1, C_2 and C_3 be the fraction of time a faculty member spends on teaching, research and management, respectively. Each C_i is quantified in four states: $c_1 = (0-25\%)$, $c_2 = (25-50\%)$, $c_3 = (50-75\%)$ and $c_4 = (75-100\%)$. The evaluation a faculty member gets in the ith area, A_i, may take values 0, 1 or 2, depending on whether the evaluation is unacceptable, acceptable or excellent, respectively. It is reasonable to assume that A_i is conditionally independent of other C_j's given C_i.

Let S represent the salary complement. The variable can take values 0, 1, 2 or 3, depending on whether the salary is decreased, maintained, increased or double increased. Then, $S = g(A_1, A_2, A_3)$ where the $g(.)$ function works as described in Table 5.2. For example, a faculty member is evaluated 2, 1 and 0 in teaching, research and management, respectively. Then, his salary will be decreased.

This example complies with ICI assumptions. In other words, the fractions of time a faculty member spends in the three areas *independently influence* the salary complement result. Thus, the only required probabilities are the $P(A_i|C_i)$, which sum up 24 probabilities (= 4 (states/cause)·3 (causes)·2 (states/$A_i - 1$)). On the contrary, without ICI assumption 384 probabilities (= 4^3 (states/combination of causes)·3 (causes)·2 (states/$A_i - 1$)) would be required corresponding to each combination of possible values of the C_i's.

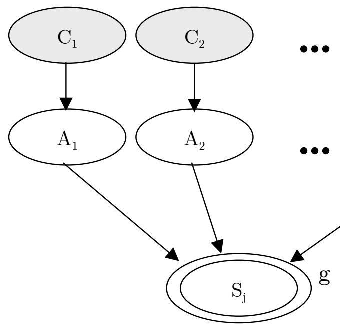


Figure 5.10: ICI model

In our application domain, modelling under ICI assumptions is carried out as follows. The intermediate variable A_i is the symptom S_j when all causes but C_i are absent ("no" or "0" state). With this definition, the conditional independence of A_i from C_h $\forall h \neq i$ is guaranteed, as required by first hypothesis of ICI model.

If $P_{S_{j,k}|C_i}$ is the probability of symptom S_j being in state $s_{j,k}$ when the cause C_i is present and the other causes are absent, then:

$$P_{S_{j,k}|C_i} = (P_{S_j} = s_{j,k}|C_i = 1, C_h = 0\ \forall h \neq i) = P(A_i = s_{j,k}|C_i = 1) \tag{5.22}$$

If the states of a variable S_i in the BN are sorted in ascending order, $k = 1, 2, ..., Q_i$, the *distinguished state* is defined as the first state, $s_{i,1}$, in that list of states. For causes with two states (no/yes), the distinguished state is the "no"(0) state. For symptoms, the distinguished state is the value of the symptom when none of the related causes are present, i.e. the "normal" state. It is assumed that when a cause C_i related to variable A_i is not present, A_i is in its distinguished state. Thus, in eq.(5.21), $P(A_i = s_{j,1}|C_i = 0) = 1$ and, consequently, $P(A_i = s_{j,k}|C_i = 0) = 0\ \forall k \neq 1$.

ICI can be extended to include situations where the symptom can be different to the distinguished state even when all causes are in its distinguished state [102]. This is very useful when there are possible causes of the symptom that are not explicitly modelled. The extension of the model consists of adding a cause C_0, named *leak cause*, whose state is always "yes", and its corresponding intermediate variable A_0, to the BN in Fig.5.10. The *background probability*, $P_{S_{j,k}|C_0} = P(A_0 = s_{j,k}|C_0 = 1) = (P_{S_j} = s_{j,k}|C_0 = 1, C_h = 0\ \forall h \neq 0)$, is defined as the probability of the symptom in absence of all the causes that are explicitly modelled. When including the leak cause in the model, the probability $P_{S_{j,k}|C_i}$ is different to $P(A_i = s_{j,k}|C_i = 1)$: $P_{S_{j,k}|C_i}$ is the probability of the symptom in the presence of only a single explicit cause, C_i, taking into account that the leak cause is also present, $P_{S_{j,k}|C_i} = (P_{S_j} = s_{j,k}|C_i = 1, C_0 = 1, C_h = 0\ \forall h \neq 0, i)$. $P(A_i = s_{j,k}|C_i = 1)$ is the probability of the symptom in the presence of only a single explicit cause, C_i, and considering that the leak cause is not present, $P(A_i = s_{j,k}|C_i = 1) = (P_{S_j} = s_{j,k}|C_i = 1, C_h = 0\ \forall h \neq i)$. The presence of the leak cause is unavoidable, thus in a knowledge-based model, experts should specify $P_{S_{j,k}|C_i}$ (see Section 5.7.1).

The intermediate variables $A_1, ..., A_K$ are not normally represented by nodes in the BN. On the contrary, they are implicit in the probability table of the S node, according to eq.(5.21). Thus, the BN structure is the one shown in Fig.5.9.

Once, the intermediate variables A_i have been defined, in order to complete the ICI modelling, the $g(.)$ function should be specified. This g function represents in which way the intermediate variables A_i, and indirectly also the causes C_i, interact to yield a common effect S_j (condition 2 on page 110). The g function can be approximated by different simple functions. Some examples are the OR function (Noisy-OR model), the maximum function (Noisy-Max model) and the addition function (Noisy-Add model). This function should be selected so that it approximates the real combined effect of the causes.

In addition, all functions should comply with: $g(A_1 = s_{j,1}, ..., A_K = s_{j,1}) = s_{j,1}$, that is, the symptom S_j is in its distinguished state, $s_{j,1}$, if all the C_i are also in their distinguished state. For example, if the causes and symptoms are binary, with states *true*/*false*, the previous assertion implies that the symptom is *false* if all causes are *false*.

In the **Noisy-OR** model [157], function g is the logical OR of its inputs:

$$g(A_1, ..., A_K) = A_1 \ OR \ A_2 \ OR \ ... \ A_K \tag{5.23}$$

This model assumes that causes and symptoms are binary ($true/false$). There exists causal mechanisms which prevent their corresponding causes from producing the symptom S_j with a probability q_i. That is, causal mechanisms are inhibitory mechanisms. Therefore, each cause C_i has a probability $P_{S_{j,2}|C_i} = P(S_j = true|C_i = true) = p_i = 1 - q_i$ of producing the symptom in the absence of all other causes.

According to eq.5.21, if the leak cause is considered, the probability of S_j given the causes can be obtained as

$$P(S_j = true|C_1, ..., C_K) = 1 - (1 - p_0) \prod_{p_i|C_i=true} \frac{1 - p_i}{1 - p_0} \tag{5.24}$$

where p_i is the probability of S_j being "true" given that cause C_i and the leak cause are "true" and all other causes are "false", and $p_0 = P_{S_{j,2}|C_0}$ is the probability of S_j given that all causes are "false", except the leak cause.

In order to build the equivalent model in Fig.5.10, the probabilities of the auxiliary nodes can be obtained as:

$$P(A_i = s_{j,2}|C_i = 1) = \frac{p_i - p_0}{1 - p_0} \tag{5.25}$$

The **Noisy-Max** model [102] is a generalization of the Noisy-OR model to deal with symptoms with multiple states. Noisy-Max assumes that the resulting "degree of manifestation" of a symptom which is related to several possible causes is the one due to the cause that has the highest impact. States should be numbered in ascending order depending on its degree of manifestation. The lowest state is the distinguished state.

In Noisy-Max model, function g is the maximum of its inputs:

$$g(A_1, ..., A_K) = max(A_1, ..., A_K) \tag{5.26}$$

When the number of states is two, the Noisy-Max model is equivalent to the Noisy-OR model. The Noisy-Max model tends to underestimated the probabilities of symptoms given several causes because the combined effect of two causes is normally higher than the effect of any of the two causes independently.

The **Noisy-Add** model [93] assumes that the effect of causes on the symptom is the sum of the effects caused by each cause independently. Therefore, function g is the sum of its inputs:

$$g(A_1, ..., A_K) = \sum_{i=1}^{K} A_i \tag{5.27}$$

Under this model, each A_i in Fig.5.10 can take on values ranging from 0 to n, whereas symptom S_j can take on values ranging from 0 to $K \cdot n$.

In order to avoid the number of states of S_j being larger than the number of states of the intermediate variables A_i, other functions can be used, such as

$$g(A_1, ..., A_K) = min(n, \sum_{i=1}^{K} A_i) \tag{5.28}$$

where n es the highest state of the intermediate nodes A_i. This model has been called **Noisy-Addlim**.

Another proposed function, which has been named **Noisy-Addceil**, is the following:

$$g(A_1, ..., A_K) = ceil(\frac{1}{K} \sum_{i=1}^{K} A_i) \tag{5.29}$$

Examples of ICI models following different g functions can be found in the Appendix 5.A to the chapter.

Model representation

The diagnosis model based on ICI assumptions is shown in Fig.5.11. In this model, each cause C_i, $i = 1..K$, is modelled as a different node, whose states are $false/true$. Contrary to the SBM and the CBM, in this model causes could have more states to represent different degrees of occurrence of the causes. However, for the sake of simplicity, only two states have been considered. In addition, one important difference with regards to the previous structures is that more than a cause can be present at the same time because causes are not mutually exclusive states of a variable.

Symptoms $S_1, ..., S_M$ are children of the causes, like in the models based on the SBM and the CBM. The main difference is that symptoms may have multiple parents, depending on the number of causes having a direct effect on that symptom. Symptoms may have any number of

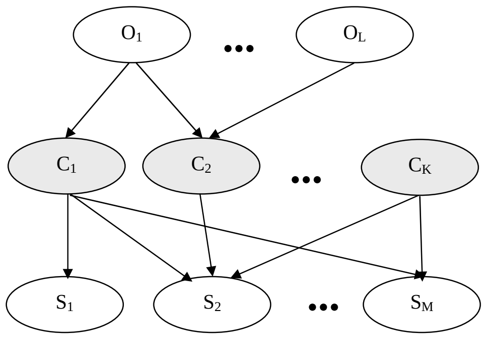


Figure 5.11: Diagnosis model under ICI assumption

states, although in order to simplify, only symptoms with two or three states have been used for the diagnosis system.

Diverse models have been built, which differ in the conditional probability tables of symptoms given causes. Those tables have been calculated based on different deterministic functions g relating the parent nodes (Noisy-OR, Noisy-Max, Noisy-Add, etc.).

Inference method

Simplified inference algorithms have been proposed for BNs based on the Noisy-OR [117, 67, 68, 208] assumption. For general BN structures, ICI has been used to transform the network structures so that the inference in the transformed BNs is more efficient than in the original networks [93, 149, 207]. Inference algorithms are out of the scope of this thesis, thus interested readers should go to the references.

5.5 Learning of model parameters

5.5.1 Methods to discretize continuous variables

Most performance indicators in cellular networks are inherently continuous. Thus, they should be discretized to build a discrete BN. In our experience of defining the diagnosis models, discretization has been found to be one of the most difficult tasks because small changes in discretization intervals lead to large variations in the diagnosis accuracy.

Most existing discretization techniques are based either on human expertise or on large training databases. A system based only on human expertise is normally very inaccurate due to the difficulties in the construction of the model. Alternatively, an adequate discretization policy could be calculated based on a large set of labelled cases. Nevertheless, if only a scarce number of training examples was available, applying classical discretization algorithms based on the available data may lead to inaccurate results. Discretization based on experts' knowledge will be further described in Section 5.7, whereas this section is focused on discretization based on data.

Discretization of continuous variables based on training data has received considerable attention in machine learning literature. This thesis is centered on univariate supervised discretization methods. *Univariate* methods are the simplest discretization algorithms, which search for the best discretization of each continuous attribute individually, without considering the relationships among attributes. *Supervised* methods utilize the class labels of the training set, in contrast to unsupervised methods, which do not make use of the classes in the discretization process. The idea of supervised discretization is to search for the partition of the value range of a continuous attribute so that its power to predict the class is maximized.

Some approaches to supervised discretization measure the class entropy with respect to the variable of interest [80, 160, 178]. Numerous studies have shown that entropy-based methods have many advantages compared to other techniques [75, 150], such as *Equal Width Interval Binning* [56] or *1R discretization* [105], thus justifying their selection for diagnosis in cellular networks.

It has been considered that one of the main figures of merit of the discretization algorithms used for diagnosis is their simplicity. This is the reason why in the following sections symptoms have only two states, i.e. $|S_i| = 2, \ \forall i$. For each continuous symptom S_i^C, the objective is to select the "best" interval limits (*thresholds*) from its range of values. Thus, the discretization problem is reduced to the definition of a single threshold, $t_{i,1} = t_i$, for each symptom S_i. Nevertheless, the proposed methods can easily be extended to more states.

In this section, the words *symptom*, *attribute* and *feature* will be used indistinctively. Likewise, the *class* will be referring to the *cause* of the problem. The *classification* is the identification of the cause. The reason for using different nomenclatures is that those terms have received different names in the literature depending on the field of application. Firstly, discretization based only on human expertise is briefly introduced. Subsequently, the basis for entropy-based methods, which rely only on data, are revised. Afterwards, an entropy-based method that combines knowledge from experts with information from training sets is proposed. The last method presented in this section finds the *maximum a posteriori* cause making use of conditional beta pdfs.

a) Experience (TEXP)

This method, which relies on knowledge only, is the most frequently used when no data is available. Domain experts, i.e. in this case experts in troubleshooting the RAN, define the thresholds. The main difficulty comes from the fact that experts are not used to deal with these parameters and, therefore, some help is required from BN engineers. Questions such as "from which value would you consider this symptom to be high?" may be asked to help experts to elicit thresholds. In addition, experience has shown that thresholds are quite dependent on the cellular network in which the diagnosis system is going to be used.

Fig.5.12 illustrates the complexity in setting thresholds. In this figure, the conditional pdfs of the symptom "Ratio of quality samples out of band 0 in uplink path" depending on diverse causes is shown. In the figure, the high degree of overlapping among the curves and the lack

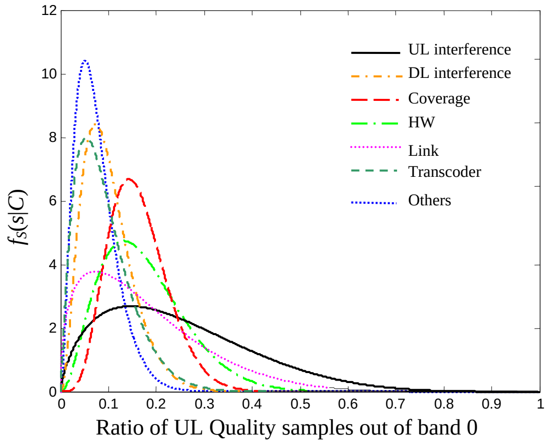


Figure 5.12: Probability density functions of quality related symptom

of clear separation limits can be appreciated. From this observation, it can be deduced that causes do not have the desirable property of clearly making the symptom increase when the cause is related to that symptom, while maintaining the symptom in a low value when the cause is not related to that symptom. For example, UL interference normally worsens the quality of the received signal. This can be noticed in the fact that the corresponding curve in Fig.5.12 is shifted to the right in comparison to the other curves. However, even if there is a problem of interference in the uplink, it can be observed that the probability of the symptom being low is also noticeable. Thus, for a human expert it is not clear where to set the threshold for this symptom.

In practice, when thresholds are elicited by experts, normally a feedback procedure of refinement is applied. It consists of testing the diagnosis system in a real network once initial thresholds have been defined, and fine-tuning them after analysis of cases where the system was erroneous in its diagnosis. This process is very time-consuming and it has to be repeated every time important features or parameters of the network are changed (such as including a new type of base station or using a new functionality).

b) Entropy Minimization Discretization (EMD)

The entropy-based method proposed in [80] will be described in this section because it is considered as state-of-the-art in discretization for classifiers.

For each continuous-valued attribute S_i^C, the algorithm is defined as follows:

1. The cases of a training set D are first sorted by increasing value of the attribute S_i^C. The midpoint m_j between the symptom values in each successive pair of cases in the sorted sequence is considered as a potential threshold.

2. Each candidate threshold m_j partitions the data set of cases D into two subsets D_1 and D_2. The class entropy of the partition is evaluated as described below.

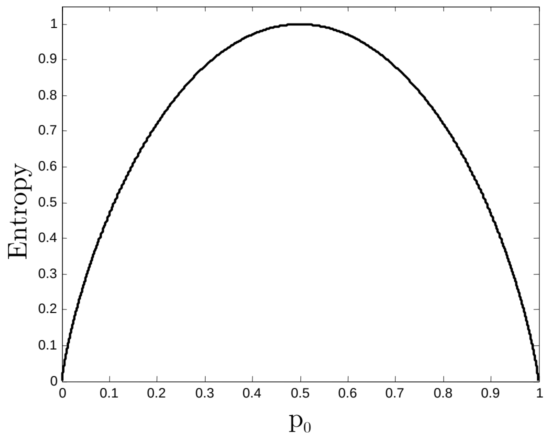


Figure 5.13: Binary class entropy

3. The best threshold t_i for S_i^C is the candidate point which minimizes the class entropy of the partition.

In [80], *boundary cut points* were defined, which are values of m_j between two cases with different classes in the sequence of sorted cases. It was proven that evaluating only the boundary cut points is sufficient for finding the minimum class entropy, which in general highly diminishes the number of candidates to be evaluated.

Let $|R|$ denote the number of cases in a subset R and let $|R(c_i)|$ be used for the number of cases in R with $C = c_i$. The class entropy of the subset R is defined as

$$Ent(R) = -\sum_{i=1}^{K} \frac{|R(c_i)|}{|R|} \cdot \log_2 \frac{|R(c_i)|}{|R|} \tag{5.30}$$

where K is the number of causes.

The utilization of the entropy can be understood from the *information theory* [170, 171]. For example, in the case of two classes, Fig.5.13 represents the entropy versus the probability of one of the classes, p_0 ($p_0 = \lim_{|R|\to\infty} \frac{|R(c_i)|}{|R|}$). It can be observed that the entropy is minimum when $p_0 = 1$ or $p_0 = 0$, i.e. when one of the two classes is certain. On the contrary, when both classes are equally possible, entropy is maximum.

The aim is to search for partitions where all cases in any of the subsets belong to the same class c_i, so that if an attribute value is within that interval, we can certainly assess that the class is c_i. Therefore, the goal of the heuristic should be to minimize the class entropy of each subset.

Let m_j be a boundary cut point of the set D of cases, which partitions it into the subsets D_1 and D_2. The class information entropy of the partition induced by m_j is the average of the class entropies of the subsets

$$Ent(D, m_j, S_i^C) = \sum_{k=1}^{2} \frac{|D_k|}{|D|} \cdot Ent(D_k) \tag{5.31}$$

where D_1 is the subset of examples in D with values of S_i^C lower than m_j and $D_2 = D \backslash D_1$. Thus, a binary discretization for D is determined by selecting the threshold t_i for which $Ent(D, m_j, S_i^C)$ is minimal among all the candidate cut points m_j.

In [80], the discretization is extended to multiple intervals by recursively applying the method explained above. A minimum description length (MDL) criterion is used to decide when to stop discretization.

c) Selective Entropy Minimization Discretization (SEMD)

A novel method, named *Selective Entropy Minimization Discretization* (SEMD), is proposed hereafter. As main contribution to the original EMD, SEMD incorporates knowledge from domain experts about the relationships between the classes and the attributes.

As described above, the goal of EMD is to select the best threshold for a continuous symptom S_i^C so that the resulting discrete symptom S_i helps to discriminate among all the classes. With this purpose, according to (5.30), EMD calculates the class entropy of a subset adding over all the classes. However, the symptom S_i^C is only related[3] to some causes, meaning that in most cases only some causes lead to an anomalous value of S_i^C, i.e. a failure. For example, the failure "excessive percentage of UL samples out of quality band 0" (Fig.5.12) normally only happens when the cause is c_1="lack of coverage", c_2 ="hardware fault" or c_3 ="UL interference", but not when the cause is any other one. Therefore, in this case, the symptom "percentage of UL samples out of quality band 0" is related to causes c_1, c_2 and c_3. It should be noticed that this is just an approximation because the pdfs of the symptom given the not related causes are different from each other. The causes related to each symptom can be found in Table 4.11.

SEMD changes the objective of EMD from differentiating among all classes to discriminating between two sets of classes: i) classes related to the symptom S_i^C, and ii) classes not related to the symptom S_i^C. The method requires the definition of the causes related to each symptom by the domain experts. Let $C_r^i = \{c_{r1}^i, ..., c_{rR_i}^i\}$ be the causes related to symptom S_i^C, and let $C_n^i = C \backslash C_r^i$ be the causes not related to symptom S_i^C. The class entropy of a subset R is now calculated as

$$Ent(R) = -\frac{|R(C_r^i)|}{|R|} \cdot \log_2 \frac{|R(C_r^i)|}{|R|} - \frac{|R(C_n^i)|}{|R|} \cdot \log_2 \frac{|R(C_n^i)|}{|R|} \tag{5.32}$$

where $|R(C_r^i)|$ denotes the number of cases in R whose class belongs to C_r^i and $|R(C_n^i)|$ are the number of cases in R whose class belongs to C_n^i.

[3]Symptom S_i^C *related* to cause c_j means that the variable S_i^C is dependent on $c_j = yes$. When the causes are represented as the states of a single cause C, the symptom S_i^C may be independent (=not related) of C only for some values of C. This is called *Asymmetric independence*.

Asymmetric independence is a situation where variables may be independent of other variables for some, but not all their values. Similarity networks [85] are one of the most important techniques to represent asymmetric independence.

The class entropy of the partition induced by a boundary cut point m_j is calculated as (5.31). The algorithm to calculate the best threshold for each symptom is the same one described on page 117, substituting (5.30) by (5.32).

The complexity of SEMD is lower than that of EMD. In (5.30) there are as many additions as number of causes, whereas in (5.32) the number of additions is two regardless of the number of causes. Thus, the number of operations of SEMD is $2/K$ times the number of operations of EMD, where K is the number of classes.

d) Beta Maximum a Posteriori (BMAP)

This novel approach, which has been named *Beta Maximum a Posteriori* (BMAP), follows directly from the theory of hypothesis testing. Statistical hypothesis testing is a formal means of distinguishing between probability distributions on the basis of random variables generated from those distributions [164].

The application of the *maximum a posteriori* probability decision rule to univariate discretization consists of selecting the class that maximizes its posterior probability given a continuous attribute S_i^C

$$\max_j [P(c_j|s_i^C)] = \max_j [f_{S_i^C}(s_i^C|c_j) \cdot P(c_j)] \tag{5.33}$$

where $f_{S_i^C}(s_i^C|c_j)$ is the conditional pdf of symptom S_i^C in s_i^C given cause c_j.

Instead of considering all causes separately, a distinction among the causes C_r^i related to a given attribute S_i^C and the causes C_n^i not directly related to S_i^C is carried out, similarly to the SEMD method. Thus, eq.5.33 becomes

$$\max_{n,r} [P(C_r^i|s_i^C), P(C_n^i|s_i^C)] \tag{5.34}$$

The threshold t_i is the cross-over point of the curves defined by (5.34), corresponding to causes related and not related to S_i^C, i.e. the value that accomplishes the following equation

$$\sum_{c_j \in C_r^i} f_{S_i^C}(t_i|c_j) \cdot P(c_j) = \sum_{c_j \in C_n^i} f_{S_i^C}(t_i|c_j) \cdot P(c_j) \tag{5.35}$$

where it has been considered that causes are exclusive and the Bayes' rule has been applied.

For example, Fig.5.14 shows both terms of the previous equality together with the cross-over point t_i for the symptom "Percentage of samples with Uplink signal level<-100 dBm". The related causes are "lack of coverage" and "hardware fault" because in the presence of any of those causes the received signal level is normally reduced.

In eq.(5.35), the prior probabilities of causes can be easily elicited by domain experts based on their experience. Thus, the main difficulty is determining the conditional pdfs of the symptoms given the causes, which are not normally known in most application domains. Nevertheless, in cellular networks the pdfs of symptoms conditioned to the causes can be accurately modelled as beta pdfs, as explained in Section 5.2.3. Hence, based on the set of training cases D, the parameters a and b can be estimated using a maximum likelihood method [164]. Alternatively, if

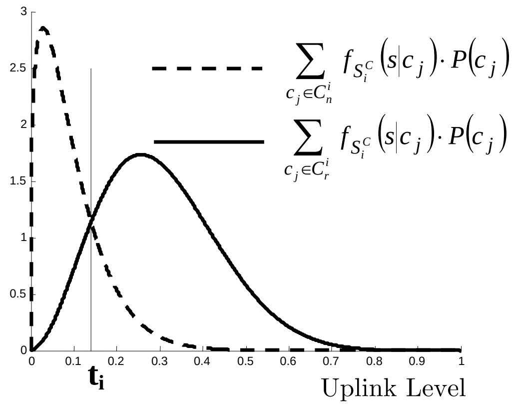


Figure 5.14: Example of BMAP discretization method

the number of available cases is reduced or non-existent, the parameters a and b can be elicited by diagnosis experts.

In summary, the BMAP discretization method is described by the following procedure:

1. Prior probabilities of causes should be determined based on data or on knowledge.

2. For each continuous symptom S_i^C and cause c_j, calculate the parameters a and b of the beta pdf that best fit the training data. Those parameters can also be elicited by experts.

3. Find the threshold t_i as the cross point between the probability of the related causes given the continuous symptom and the probability of the non-related causes given the continuous symptom.

5.5.2 Methods to estimate probabilities

The aim of this section is to present different alternatives for defining the probabilities of a discrete BN. In particular, due to the structures of the BNs presented in Section 5.4, we are interested in the prior probabilities of the conditions $P(o_{k,h})$, the conditional probabilities of causes given conditions $P(c_i|o_{k,h})$ and the conditional probabilities of the discretized symptoms given the causes $P(s_{j,g}|c_i), \forall i, j, k, h, g$. In order to simplify the model, it will be assumed that symptom S_j has been discretized into only two states $s_{j,1}$ and $s_{j,2}$. In the following sections $P(S_j = s_{j,1}|c_i)$ will be expressed as $P(s_{j,1}|c_i)$. Furthermore, only $s_{j,1}$ will be computed because $P(s_{j,2}|c_i) = 1 - P(s_{j,1}|c_i)$.

Firstly, probability elicitation based only on human expertise is summarized. The described method in the following section, Maximum Likelihood Estimation (MLE), calculates the probabilities of the BN based on relative frequencies of occurrence in training data. The following algorithm, m-estimate (MEST), assumes beta prior probabilities to compute the probabilities. Finally, last section presents a method, Beta Distribution Function (BDF), to calculate the probabilities from beta distribution functions obtained from a training set.

a) Experience (EXP)

The first method is based on knowledge only. Experts in diagnosis elicit the probabilities of the BN based on their experience. Several techniques [175, 163, 188] have been proposed to obtain those numbers from domain experts who are normally reluctance to provide numerical probabilities.

In our experiments (see Section 5.7 for more details about knowledge acquisition), experts were asked to choose one out of five levels for the probabilities: Almost certain, Likely, Fifty-Fifty, Improbable, Unlikely. Those levels were mapped to the following probabilities: 0.85, 0.7, 0.5, 0.3, 0.1, respectively.

b) Maximum Likelihood Estimation (MLE)

MLE is one of the most commonly used estimators in statistics. When applied to probability estimation, it approximates the probabilities with relative frequencies

$$P(s_{j,1}|c_i) = \frac{|D(c_i, s_{j,1})|}{|D(c_i)|} \;;\; P(c_i|o_{k,h}) = \frac{|D(c_i, o_{k,h})|}{|D(o_{k,h})|} \;;\; P(o_{k,h}) = \frac{|D(o_{k,h})|}{|D|} \tag{5.36}$$

where

- $|D(o_{k,h})|$ is the number of cases in D where the state $o_{k,h}$ of the condition O_k is observed
- $|D(c_i)|$ is the number of cases in D where the cause c_i is observed
- $|D(c_i, o_{k,h})|$ is the number of cases in D where both the cause c_i and the state $o_{k,h}$ of the condition O_k is observed
- $|D(c_i, s_{j,1})|$ is the number of cases in D where both the cause c_i and the state $s_{j,1}$ for the symptom S_j are observed.

The problems with this approach arise when $|D(o_{k,h})|$, $|D(c_i, o_{k,h})|$, $|D(c_i, s_{j,1})|$ or $|D(c_i)|$ are low or even zero because the estimated probabilities will be inaccurate. Unfortunately, this is normally the situation when the number of cases is scarce.

c) M-estimate (MEST)

The MEST method overcomes the poor results obtained with MLE when some probabilities are small and the number of training cases is scarce.

Firstly, in order to reduce the problems when either $|D(o_{k,h})|$ or $|D|$ are small, Laplace's law of succession [86, 114] is applied. Thus, the prior probability of a condition $o_{k,h}$ is estimated by the probability of the condition happening in the next trial when there were $|D(o_{k,h})|$ cases where the condition was $o_{k,h}$ in the $|D|$ previous cases. This rule assumes that the initial distribution of conditions is uniform. Thus, the probability of the condition $o_{k,h}$ is estimated as

$$P(o_{k,h}) = \frac{|D(o_{k,h})| + 1}{|D| + W_k} \tag{5.37}$$

where W_k is the number of states of condition O_k.

Secondly, for the estimation of conditional probabilities, beta pdfs are preferred to uniform pdfs as the initial probability distributions. Accordingly, the conditional probability of causes given conditions $P(c_i|o_{k,h})$ can be computes using the *m-estimate* [140, 58, 201] as

$$P(c_i|o_{k,h}) = \frac{|D(c_i, o_{k,h})| + m \cdot P(c_i)}{|D(o_{k,h})| + m} \tag{5.38}$$

where $P(c_i)$ is estimated by Laplace's law of succession as

$$P(c_i) = \frac{|D(c_i)| + 1}{|D| + K} \tag{5.39}$$

where K is the number of causes.

Likewise, the conditional probability of symptoms given causes $P(s_{j,1}|c_i)$ can be calculated as

$$P(s_{j,1}|c_i) = \frac{|D(c_i, s_{j,1})| + m \cdot P(s_{j,1})}{|D(c_i)| + m} \tag{5.40}$$

where $P(s_{j,1})$ is estimated by Laplace's law of succession as

$$P(s_{j,1}) = \frac{|D(s_{j,1})| + 1}{|D| + 2} \tag{5.41}$$

The parameter m is a constant related to the a and b parameters of the beta pdfs. In our experiments, the default value $m = 2$ has been used because it empirically gives good results, according to [58]. Alternatively, m could be estimated based on the parameters a and b of the prior beta pdf estimated by experts, according to the following. Let β be a random variable following a beta distribution with parameters a and b. The average prior probability of β can be calculated (eq.5.9) as

$$E_{prior}(\beta) = \frac{a}{a + b} \tag{5.42}$$

In Section 5.2.2, it was explained that if the prior probability of a random variable β was beta with parameters a and b, the posterior probability after n successes and $N - n$ failures were observed in N trials was also beta with parameters $a + n$ and $b + N - n$. Therefore, the average posterior probability can be obtained as

$$E_{posterior}(\beta) = \frac{a + n}{a + b + N} = \frac{m \cdot E_{prior}(\beta) + n}{m + N} \tag{5.43}$$

where $a + b$ has been called m. Eq.5.38 and 5.40 are particularizations of eq.5.43.

Therefore, experts in diagnosis the RAN could estimate the parameters a and b of the prior distributions for each conditional pdf and, then, m could be calculated as $a + b$ and E_{prior} could be computed following eq.5.42. Instead of a and b, other parameters easier to estimate by expert could be defined, e.g. the mean and standard deviation. Alternatively, experts could directly specify E_{prior} and m. In that case, m gives an idea about the standard deviation of the beta distribution: if the experts feel confident about asserted E_{prior}, that means that variance is small and m should be large; on the contrary, if the expert is not sure of E_{prior}, variance is large and m should be small.

According to (5.37)-(5.41), it can be noticed that even if $|D(o_{k,h})|$, $|D(c_i, o_{k,h})|$, $|D(c_i, s_{j,1})|$, $|D(c_i)|$ or $|D|$ are zero, the estimated probabilities will be different from zero.

d) Beta Distribution Function (BDF)

The BDF method is used to estimate the conditional probabilities for symptoms, which are the most complex probabilities to define. In that case, the prior probabilities of conditions and the probabilities of causes given conditions can be elicited by experts or calculated using the Laplace or m-estimate algorithms.

The BDF method consists in calculating the conditional probability of the first state of the symptom as the value of the distribution function in the threshold between the two states.

Firstly, training data should be fitted to beta pdfs based on expertise or using a maximum likelihood method [164]. Once the parameters of the beta pdfs have been estimated for all symptoms, the conditional probabilities can be calculated as the area under the beta curve in each interval

$$P(s_{j,1}|c_i) = F_{S_j^C|c_i}(t_j) \tag{5.44}$$

where $F_{S_j^C|c_i}(s)$ is the distribution function of the symptom S_j^C conditioned to cause c_i. $F_{S_j^C|c_i}(s)$ is a beta distribution function (i.e. the integral of a beta pdf) with parameters a and b calculated based on the training data or defined by experts. The threshold t_j can be calculated applying any of the methods described in Section 5.5.1.

For example, Fig.5.15 shows the conditional pdf for the symptom "Percentage of samples with Uplink signal level<-100 dBm", together with the threshold calculated according to BMAP. The shadowed area (0.075) is the probability of the symptom being in the first state (low) conditioned to a "HW fault". Thus, the probability of the symptom being in the second state is 0.925.

5.6 Prevention of imprecision in model parameters

5.6.1 Introduction

As explained before, in cellular networks normally there are no databases of classified training examples. Due to the difficulty to get training cases, very often the only option from the ones presented in section 5.5 is to define the parameters of the model using the experience of RAN troubleshooters.

In the discretization of continuous symptoms, the number of states should be kept as low as possible in order to ease the knowledge acquisition process. However, if the number of states is too low, the accuracy of the diagnosis decreases because the approximation to the real behavior is coarse. For example, Fig.5.16 shows the posterior probability of a cause given the available

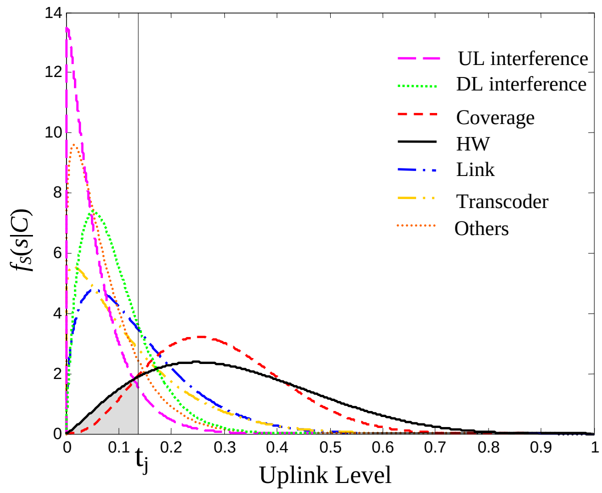


Figure 5.15: Example of probability calculation according to BDF method

evidence vs. the value of a continuous symptom S. The dotted lines are the results obtained from discrete BNs where the continuous variables were discretized (either into 2 or 3 states), whereas the continuous line represents the posterior probability obtained from a Bayesian classifier where the continuous symptoms were not discretized, therefore its behavior is closer to reality. In this example, when the symptom was discretized into two states, the threshold was set to 19%. That means that if the symptom was lower than 19% it was considered low. In that case, the output of the diagnosis system for the cause DL interference was a probability of 0.1. On the contrary, if the symptom was higher than the threshold, the probability was 0.85. These values are obtained regardless of how close to the threshold the value of the symptom is, e.g. values of the symptoms of 18.9% and 19.1% would lead to quite different conclusions in the diagnosis results. This behaviour is due to the steep shape of the posterior probability, which is far from the real statistical behaviour and also far from the way of thinking of a human expert. Increasing the number of states to solve the previous problem has been considered unfeasible because even specifying thresholds and probabilities for three states per symptom was considered too demanding by experts.

The objective of this section is to propose novel diagnosis systems which improve the diagnosis accuracy without increasing the complexity of knowledge acquisition [40]. The aim of the new methods is obtaining a posterior probability of causes closer to the real behavior. The data supplied by diagnosis experts should be the same as the ones defined for discrete BNs with 2-state symptoms. With this purpose, the first technique, which has been called Smooth Bayesian Networks (SBN), tries to smooth the steep shape of the posterior probability of the causes around the threshold. In the second method, which has been named Multiple Uniform Intervals (MUI), the BN is composed of 3-state symptoms, although the expert only have to specify a single threshold per continuous symptom. Both techniques, SBNs and MUIs, can be used not only onto a knowledge based BN, but also when the BN has been obtained based

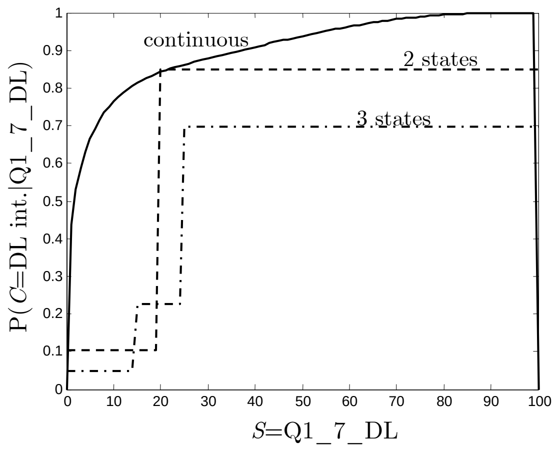


Figure 5.16: Posterior probability of the cause "DL interference" vs. the value of a continuous symptom S

on training data. Nevertheless these methods are expected to be beneficial specially when the parameters in the initial BN are inaccurate, so they are recommended to be used in combination with knowledge-based models.

5.6.2 Smooth Bayesian Networks (SBN)

Definition of SBNs

A SBN is a discrete BN with uncertainty about the states of each variable. This is the reason why in SBNs *likelihood evidence* is set for the symptoms, instead of the traditional evidence. Evidence normally determines that a symptom is in a single established state. On the contrary, likelihood evidence (also called *virtual evidence*) revises a probability distribution under uncertain evidence, i.e. there is no certainty about which of the mutually exclusive states is taken by a given variable [157, 59, 187]. SBNs can also be understood as a means of equalizing the probabilities of the states of a variable. In the limit, when there is no certainty at all about the parameters (thresholds or probabilities), the probabilities of all states should be the same.

SBNs smooth the posterior probabilities of the causes given the values of the symptoms in order to approximate them to the reasoning mechanisms applied by human experts. With that purpose, SBNs estimate the conditional probabilities of the symptoms given the causes by using simple distribution functions, as it will be explained below.

A BN equivalent to a SBN can be built as follows. Let B be a discrete BN. A continuous child, R_i, is added to each symptom, S_i, that is continuous in its origin (Fig.5.17). The continuous symptom related to the discrete symptom S_i will be represented as S_i^C. The pdfs $f_{R_i|S_i}(R_i = s|S_i = s_{i,j})$, $j = 1, ..., Q_i$, will be called *belief mapping functions*, $f_j^{S_i}(s)$, because they model the belief in the value of S_i^C given that S_i is in the state $s_{i,j}$. For example, let's consider the continuous symptom S_i^C ="Percentage of handovers due to interference" and let's assume that

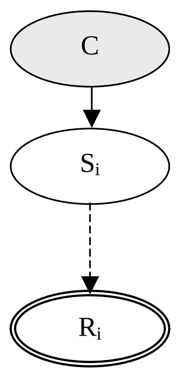


Figure 5.17: Addition of a child to the discrete symptom S in order to "smooth" the BN

the discrete symptom S_i has three states: low / medium / high. If the value of S_i^C is $s = 20\%$, $f_{R_i|S_i}(R_i = 20\%|S_i = low)$ represents the pdf of the continuous symptom in the value 20%, knowing that this value is considered low by an expert. Most existing commercial tools for BNs include functionalities to deal with virtual evidence, without the need to implicitly adding the continuous nodes, R_i.

A SBN can be defined by the original BN B, the specification of the nodes to be smoothed, $S_R = \{S_1, ..., S_W\}$, and the set $F = f_j^{S_i}(s)$, $i = 1, ..., W$, $j = 1, ..., Q_i$, of belief mapping functions. According to the laws of probability, the belief mapping functions should comply with:

$$\int_{S_i^C} f_j^{S_i}(s)\, ds = 1 \quad \forall i, j \tag{5.45}$$

It can be easily demonstrated that SBNs approximate the probability of a continuous symptom S_i^C, given a cause C_k, as linear combinations of the belief mapping functions in s:

$$\hat{f}_{S_i^C|C_k}(s|\, C_k) = \sum_{j=1}^{Q_i} p(S_i = s_{i,j}|\, C_k) \cdot f_j^{S_i}(s) = \sum_{j=1}^{Q_i} a_{i,j}^k \cdot f_j^{S_i}(s) \tag{5.46}$$

where the $a_{i,j}^k$ depend on the probabilities obtained from the original discrete BN B.

A SBN equivalent to the initial discrete BN B can be built using rectangular belief mapping functions (Fig.5.18(a)), meaning that any value between the boundaries of a given state is equally probable. Nevertheless, a human expert's way of reasoning is far from that behavior. For example, knowing that a symptom is low, an expert would think that low values of S_i^C are more probable. However, this probability would not be constant for all values lower than the threshold (like it is the case when using rectangular functions), but it would gradually decrease when approaching the threshold.

Posterior probabilities

It can be demonstrated (5.B) that the posterior probability of a cause C_k given the values of the continuous symptoms S_i^C can be obtained using the following equation:

$$P\left(C_k \middle| s_1 s_2 \ldots s_M\right) = \frac{\sum\limits_{S_1 S_2 \ldots S_M} P(C_k, S_1, ..., S_M) \cdot \prod\limits_{h=1}^{M} f^{S_h}(s_h)}{\sum\limits_{S_1 S_2 \ldots S_M} P(S_1, ..., S_M) \cdot \prod\limits_{h=1}^{M} f^{S_h}(s_h)} \tag{5.47}$$

Belief mapping functions

In the following pages some examples of belief mapping functions will be presented. In order to make the knowledge acquisition affordable for the experts only two states per symptom will be used to discretize the continuous variables. Complex functions and more states may increase the diagnosis accuracy, but also increase the need for detailed input from domain experts.

The *transition zone* of a set of belief mapping function $f_j^{S_i}$, $j = 1, 2$, is defined as a range of values centered in the threshold $T = t_{i,1}$ and having a width of p (which is called *degree of smoothness*). The threshold T is also the point where the belief mapping functions decay to half of their maximum value.

There is a trade-off in selecting the adequate belief mapping functions: on the one hand, achieving high classification accuracy, and, on the other hand, simplifying the knowledge acquisition process. Firstly, the belief mapping functions should be selected so that (5.46) is as close as possible to the actual statistical behavior of the continuous variable.

Secondly, the number of parameters for the functions should be low. It is recommended that the parameter p of the mapping functions is set as default by the model designer and it is not required that the diagnosis experts change it, in order not to increase the information provided during knowledge acquisition. At the most, the expert could be asked for his/her level of confidence in the defined parameters. If the expert is confident with the parameters in the model, p should be low. Otherwise, p should be medium or large (e.g. we have empirically found out that $p = 20\%$ is normally a good selection when parameters are not very accurate).

A third consideration is that the diagnosis should not be very sensitive to small changes in the parameters of the mapping functions, in particular to the thresholds. For example, in the case of traditional BNs, a small change in the definition of the thresholds may lead to different conclusions in the diagnosis, especially if the values of the symptoms are near the thresholds. The sensitivity to changes in a given parameter x can be measured as

$$\frac{\partial P\left(C_k \middle| E\right)(x)}{\partial x} \tag{5.48}$$

where $P(C_k|E)$ is the posterior probability of the cause C_k given the available evidence E and x is a parameter of a belief mapping function.

In this thesis, some simple belief mapping functions are proposed (see 5.C for the related equations):

- **Rectangular functions (rect)**

  The original discrete BN B is equivalent to a SBN with rectangular mapping functions. The rectangular belief mapping functions for a continuous symptom S_i^C discretized in two states are shown in Fig.5.18(a). They are uniform pdf in the interval corresponding to a given state.

  This set of mapping functions depends on a single parameter: the threshold between the states, T. In this case, the degree of smoothness p is always zero. The main advantage of the rectangular functions is that the knowledge acquisition and the BN implementation are very simple. However, the sensitivity is very high for values of the evidence around the threshold T.

- **Trapezoidal functions (trap)**

  The trapezoidal mapping functions for a continuous symptom S_i^C discretized into two states are lines in the transition zone, as shown in Fig.5.18(b). These functions are aimed to reduce the slope of the rectangular functions around the threshold.

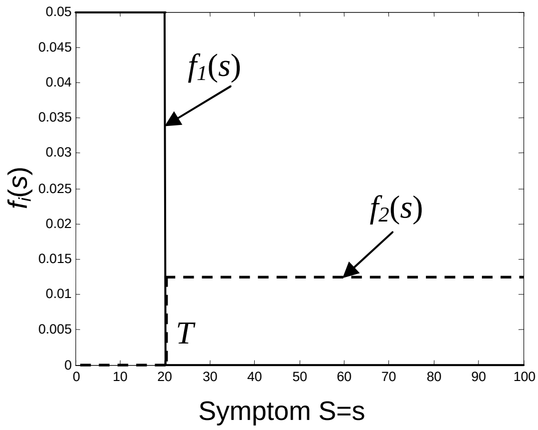
  This set of functions depends on two parameters: the threshold T and the degree of smoothness p. The parameter p has been called degree of smoothness because it is related to the slope of the functions. In this case, the slope is different from zero only in the transition zone. It should be pointed out that the functions are normalized to comply with (5.45). Furthermore, p is limited so that the transition zone lies within the range of the symptom.

- **Rect-Gaussian functions (gaus)**

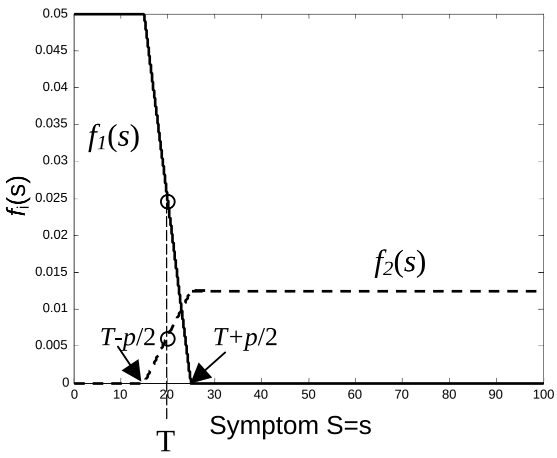
  These types of mapping functions (Fig. 5.18(c)) are related to Gaussian distributions. Similarly to the trapezoidal functions, they depend on two parameters: the threshold T and the degree of smoothness p. The function corresponding to the first state is constant for values below $T-p/2$ and a gaussian function, whose average is $T-p/2$, from that value onwards. Likewise, the function corresponding to the second state is a gaussian function, whose average is $T+p/2$, for values below $T+p/2$ and constant above that value. Thus, the degree of smoothness p is related to the standard deviation of the Gaussian distribution according to $p = 2\sigma\sqrt{ln\ 4}$. Similarly to trapezoidal functions, rect-gaussian functions are normalized and p is limited.

Related work

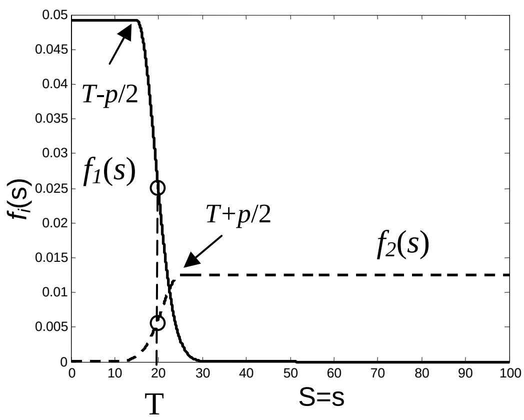
Several authors have tried to incorporate the continuity of human reasoning in BNs. In [152, 153, 151] two different components in uncertainty are distinguished: uncertainty in the output of a clearly defined and randomly occurring event (described by a probability) and uncertainty inherent in the description of the event itself (described by fuzzy logic). Thus, in those papers, an integration of BNs and fuzzy logic [206], called *Fuzzy Causal Probabilistic Networks* (FCPN) is proposed. In a FCPN there is a fuzzifier and a defuzzifier associated to each discrete variable



(a) Rectangular belief mapping function



(b) Trapezoidal belief mapping function



(c) Rect-Gaussian belief mapping function

Figure 5.18: Types of belief mapping functions

of a BN. A fuzzifier converts a continuous variable to a fuzzy variable, whereas a defuzzifier does the inverse. In [151] Conditional Gaussian models are used for mapping between continuous variables and their partner fuzzy variables. In [200] a methodology is presented to estimate the pdf of a continuous variable from the probabilities of the discrete variables in the BN. Using the estimated pdfs and the prior probabilities, the posterior probability of the hypothesis given the continuous evidence can be determined. The estimated density functions depend on the definition of the membership functions, which are characteristics of fuzzy sets. [159] uses two approaches to diagnose print defects: BNs and fuzzy logic. The final diagnosis is a combination of the conclusions of the two systems. Finally, [60] uses fuzzy logical operators and possibility vectors to design a *fuzzy belief network*.

Amongst previous references, the technique in [152] is the closest to the one proposed in this thesis, SBNs. The main differences between these two are:

- Reference [152] is directly related to the theory of fuzzy logic. SBNs are only based on probability theory.

- Reference [152] uses *soft evidence*, whereas SBNs apply *likelihood evidence*. Soft evidence specify the probability distribution of a variable. Likelihood evidence revises a probability distribution under uncertain evidence. More information about the differences between soft evidence and virtual evidence can be found in [187].

- The methods presented in references [152, 151, 200] achieve the same ultimate results as SBNs: smoothing the posterior probabilities of the hypothesis given the value of continuous variables. However, the procedures followed to get the smoothing are different and so it is the actual shape of the posterior probability.

Although the problem of handling continuous variables has already been tackled in the literature, none of the papers described above quantified the benefits of using the proposed methods instead of the traditional discrete BNs. Assessing the effects of using different functions to approximate the continuous pdf is another aspect that has not been previously described.

5.6.3 Multiple Uniform Intervals (MUI)

A straightforward approach to improve the diagnosis performance of the diagnosis systems based on BNs is increasing the number of intervals in which the continuous symptoms are discretized. The drawback is that the knowledge acquisition is complicated because the number of parameters increases significantly. The algorithm proposed in this section, which will be called Multiple Uniform Intervals, increases the number of intervals while the amount of information required from experts remains the same compared to the case when only two intervals are used [40].

Let S_i^C be a continuous symptom, which has been discretized into two intervals and let S_i be the resulting discrete symptom. The parameters elicited by the experts are the probability of the first state given each cause C_k, $p_{i,1}^k$, and the boundary between the two states, T.

$$p_{i,1}^{k} = P\left(S_i = s_{i,1} \middle| C_k\right) \tag{5.49}$$

$$S_i = s_{i,1} \quad \text{if } S_i^C \leq T, \quad \text{else} \quad S_i = s_{i,2} \tag{5.50}$$

The algorithm consists of creating a third interval centered in T, whose width p is a parameter equivalent to the degree of smoothness characteristic of SBNs. In order to set the probability of the new state, it is assumed that the continuous symptom given the cause follows a uniform distribution. The uniform distribution has been chosen because it is the most adequate distribution when there is no prior certainty. Thus, the probability of the second state is:

$$p_{i,2}^{k*} = \frac{p}{b-a} \tag{5.51}$$

where a and b are the lower and upper limits, respectively, of the continuous symptom.

The probabilities of the other two states have to be updated to

$$p_{i,1}^{k*} = p_{i,1}^{k} \cdot \left(1 - \tfrac{p}{b-a}\right);\ p_{i,3}^{k*} = \left(1 - p_{i,1}^{k}\right) \cdot \left(1 - \tfrac{p}{b-a}\right) \tag{5.52}$$

For the SBM it can be easily demonstrated that the posterior probability of the cause C_i given the symptoms is:

$$P(C_i|S_1, ..., S_M) = P(C_i) \cdot \frac{\prod\limits_{S_j \notin S^2} P(S_j|C_i)}{\sum\limits_{k} P(C_k) \prod\limits_{S_j \notin S^2} P(S_j|C_k)} \tag{5.53}$$

where S^2 is the set of symptoms S_i whose state is the second one, $S_i \in S^2 \Leftrightarrow S_i = s_{i,2}$.

It should be pointed out that the posterior probabilities of the causes obtained with eq.(5.53) are independent on the assumptions taken for the pdf of the symptoms (i.e. uniform or any other pdf). This is due to the fact that probabilities of the first and last state are updated by multiplying the initial probability by $1 - p_{i,2}^{k*}$ (see eq.5.52). Consequently, the same factors appear in the numerator and denominator of eq.(5.53).

5.7 Knowledge Acquisition

A BN consists of a qualitative and a quantitative part. The qualitative part is composed of nodes and arcs among them, that is the variables in the model and their dependencies. The quantitative part are the probabilities, which link the variables. Probabilistic information can be obtained from diverse sources. The most common ones are statistical data, literature and human experts [77]. Firstly, in many application domains, such as medical diagnosis, large data collections are available [50] documenting previously solved problems. These data can be used to automatically build the BN structure and to calculate the quantitative part that best fits the available information [143, 53, 64, 94]. Secondly, literature often provides probabilistic information in some application domains. However, this information is usually not directly applicable to model construction because of diverse reasons: not all probabilities are provided, probabilities are expressed in a direction reverse to the direction required by the BN, population from which information is derived is different from the population for which the BN is being developed, etc. Finally, when there are few or no reliable data available, the knowledge and experience of experts in the domain of application is the only source of information to build the BN.

If the diagnosis model is based on discrete BNs, quantitative information should also include the discretization of continuous variables. As this aspect is something external to the BN, literature related to construction of BNs normally does not mention this important part of the model design. Due to the fact that in mobile communications networks most symptoms are inherently continuous, in this thesis discretization has been considered a crucial issue in the definition of the quantitative model. Similarly to probabilities elicitation, discretization can also be based on data, literature or expertise.

In mobile communication networks, currently there are not historical collections of diagnosed cases. Furthermore, diagnosis of the RAN of cellular networks is not documented in existing literature. Thus, experience of troubleshooting experts is, in most cases, the only source of information to build a diagnosis model. Definition of the quantitative model based on statistical data was presented in Sections 5.5.1 and 5.5.2. This section is focused on knowledge acquisition.

Knowledge acquisition (KA),i.e. building a BN from the knowledge of experts in the application domain, involves two phases. Firstly, *knowledge gathering*, that is obtaining the knowledge from experts. Secondly *model construction*, that is defining the model based on the previously acquired information. KA has been considered the bottleneck of BNs because the parameters (e.g. number of probabilities) involved in a large BN is normally intractable to be specified by human experts. Hence, KA requires a trade-off between a large and detailed model to obtain accurate results on the one hand, while, on the other hand, keeping the cost of construction and maintenance and the complexity of probabilistic inference at a reasonable level. In KA, several problems are often encountered. On the one hand, experts in the application domain are not normally used to the terminology used in BNs. In addition, experts feel reluctance to specify precise quantitative information. On the other hand, experts' time is scarce, while KA is normally a very time-consuming task. Therefore, several techniques have been proposed to simplify knowledge acquisition [175, 188, 76].

Based on the theory presented in the following sections, a tool may be built in order to automatically carry out knowledge acquisition [36, 42], which has been named *Knowledge Acquisition Tool* (KAT). KAT is envisaged to guide the expert through a sequence of questions regarding his/her way of reasoning in diagnosis. A diagnosis model is automatically constructed based on his/her answers. The main advantage of KAT is that it is very easy to use by troubleshooting experts as no BN knowledge is required to use the tool. As a consequence, domain experts can transfer their expertise using a language that they understand. It should be taken into account that model construction depends on the BN structure. Therefore, user should specify which type of model he/she wishes to build.

In Section 5.7.1, KA is presented when the target is to build a discrete BN [42]. Likewise,

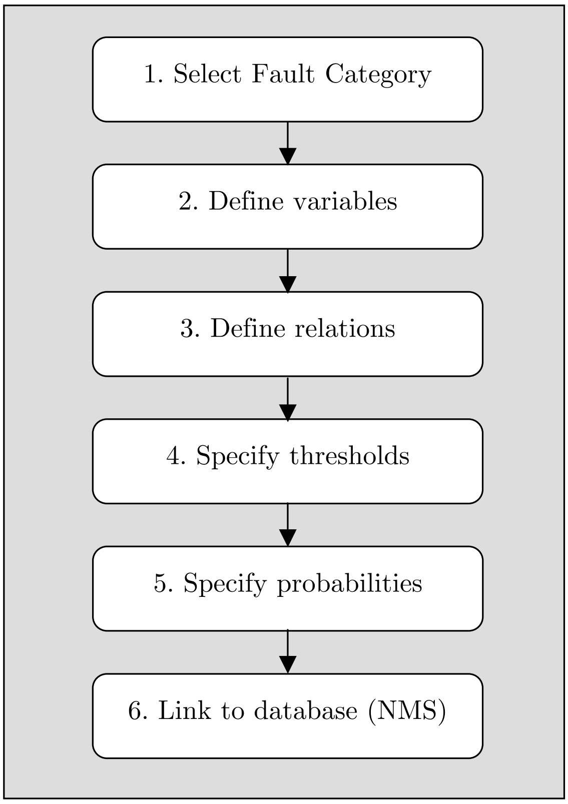


Figure 5.19: Phases in Knowledge Acquisition for BN

Section 5.7.2 presents KA for a diagnosis model based on a BC.

5.7.1 Bayesian Network

Knowledge gathering

Knowledge gathering is composed of the phases presented in Fig.5.19, which will be explained in the paragraphs below. Table 5.3 summarizes the qualitative information that the expert should provide, whereas quantitative information can be found in Table 5.4.

1. **Select *Fault Category*.** Fault categories are the diverse problems that the RAN may suffer, such as "High DCR" or "Congestion". A different model is built for each fault category. Although this thesis is focused on the problem "High DCR", the proposed KA is also valid to construct models for other fault categories.

2. **Define variables.** There should be a database of causes, symptoms and conditions. The expert has the chance of either selecting a variable from the database or defining a new one, which should then be incorporated into the database. If the number of variables in the database is large, it may be very time-consuming to read all of them in order to find a cause similar to the one that the expert wants to define [167]. In that case, once the user

has described the variable, KAT should find and present similar variables, e.g. the terms "HW fault" and "fault in a piece of equipment" should be merged in the search.

Firstly, the expert specifies the possible causes of the fault category, that is the causes of the problem in the network for which the diagnosis model is being built (e.g. "High DCR"), $\{C_1, ..., C_K\}$. It is recommended to include a cause called "Other causes", in order to cover any other possible cause of the problem not explicitly included in the defined causes. Secondly, the expert is demanded to enumerate the symptoms that may help to identify the previously defined causes, $\{S_1, ..., S_M\}$. The states, $s_{i,j}$, of each symptom, S_i, should also be specified. Lastly, the user is requested about conditions, $\{O_1, ..., O_L\}$, and their states, $o_{i,j}$, which may also help to identify the cause.

3. **Define relations**. In this phase, the user should define which are the causes, $C_r^i = \{C_{r1}^i, ..., C_{rR_i}^i\}$, related to each symptom S_i. The term "related" is used to qualify those variables which have a strong direct inter-dependency (see page 119). For example, the cause "Lack of coverage" is related to the symptom "Percentage of UL samples with level < -100 dBm", whereas the cause "UL interference" is not related to that symptom. The explanation is that a lack of coverage reduces the received signal level in comparison to the average received signal level in a network without problems, whereas when the cause is interference, the received signal level is not significantly decreased in comparison to the level in a cell without problems. The causes not related to symptom S_i will be denoted $C_n^i = C \backslash C_r^i$.

   The expert should also specify which are the conditions, $O_r^i = \{O_{r1}^i, ..., O_{rU_i}^i\}$, associated to each cause C_i, that is conditions whose value can modify his/her belief in the probability of the cause being the one causing the problem.

4. **Specify thresholds**. For each continuous symptom, S_i, interval limits (i.e. thresholds), $t_{i,j}$, between each defined interval should be requested from the user.

5. **Specify probabilities**. Verbal probability expressions are often suggested as a method of eliciting probabilistic information [163]. The number of verbal expressions should be reduced in order to avoid misinterpretations. In addition, it is advisable to use a graphical scale with numbers on the one side and words on the other. In our experiments with cellular network operators, experts were asked to choose one out of five levels of probabilities: "Almost certain", "Likely", "Fifty-fifty", "Improbable" and "Unlikely". Those levels are mapped to the probabilities 0.85, 0.7, 0.5, 0.3 and 0.1, respectively.

   The procedure to define the probabilities is as follows. Firstly, the expert is requested about the prior probabilities of each of the possible causes of the problem, P_{C_i}. As causes have only two states (no/yes), only the probability of the cause being present is demanded. In the case of a cause C_i related to a condition O_j, probability of C_i should be defined for each state of O_j. If more than a condition is related to C_i, probability of C_i should be defined for each combination of states of the associated conditions, $P_{C_i|O_i^r}$. Very often,

Table 5.3: Qualitative model defined by expert

Parameters	**Range**	**Description**	**Example**
F_i	$i = 1, ..., A$	Fault categories A: number of fault categories	F_1 =High DCR
C_i	$i = 1, ..., K$	Causes K: number of causes	C_1 =UL interf.
S_i	$i = 1, ..., M$	Symptoms M: number of symptoms	S_{30} =% UL interf.HOs
$s_{i,j}$	$i = 1, ..., M$ $j = 1, ..., Q_i$	Symptom states Q_i: number of states of symptom S_i	$s_{30,1}$ =low
O_i	$i = 1, ..., L$	Conditions L: number of conditions	O_2 =Frequency Hopping
$o_{i,j}$	$i = 1, ..., L$ $j = 1, ..., W_i$	Condition states W_i: number of states of condition O_i	$o_{2,2}$ =on
C_r^i $= \{C_{r_1}^i, ..., C_{r_{R_i}}^i\}$	$i = 1, ..., M$	Set of causes related to symptom S_i R_i: number of causes related to S_i	$C_r^1 = \{C_3, C_4\}$
O_r^i $= \{O_{r_1}^i, ..., O_{r_{U_i}}^i\}$	$i = 1, ..., L$	Set of conditions related to cause C_i U_i: number of conditions related to C_i	$O_r^1 = \{O_2\}$

only some combinations of states are implemented in the network, thus the expert should have the option of defining only those combinations that make sense. The probabilities for impossible combinations of conditions should be set to zero. If the number of conditions is large, the number of probabilities to be defined may become intractable. However, experience with cellular network operators has shown that the number of defined conditions is normally kept low, and so is the number of demanded probabilities.

The second step is defining prior probabilities of conditions, $P_{O_{i,j}}$. The number of probabilities to be specified for each condition depends on its number of states, W_i.

Lastly, probabilities for symptoms are requested. For a symptom S_i, KAT should ask the probability of each state of the symptom given that each of the related causes, $C_k \in C_r^i$, is present and the other causes are absent, $P_{S_{i,j}|C_k}$. In addition, the probability of each state of the symptom given that none of the related causes are present should be defined, $P_{S_{i,j}|C_0}$.

In all cases, the expert should take into account that the sum of the probabilities over all the states of a given symptom or condition should sum up 1. KAT should warn the expert if this is not the case.

6. **Link symptoms and conditions to database**. The last step is linking the variables in the model to the data in the NMS. Thus, symptoms and conditions should be related to a parameter (performance indicator, counter, etc.) available in the NMS or a combination of parameters. For this last option, KAT should ease the construction of equations.

Table 5.4: Quantitative model defined by expert

Parameters	Range	Description	N^{er} parameters
$t_{i,j}$	$i = 1, ..., M$ $j = 1, ..., T_i$	Threshold j for symptom S_i T_i: number of thresholds of symptom S_i	$\sum_{i=1}^{M} T_i$
$P_{C_i\|O_r^i}$	$i = 1, ..., K$	Probability of cause $C_i = on$ given set of related conditions	$\sum_{i=1}^{K} \prod_{j=r_1}^{r_{U_i}} W_j$
$P_{O_{i,j}}$	$i = 1, ..., L$ $j = 1, ..., W_i$	Prior probabilities of conditions	$\sum_{i=1}^{L} W_i$
$P_{S_{i,j}\|C_k}$ $\forall C_k \in C_r^i$	$i = 1, ..., M$ $j = 1, ..., Q_i$	Probability of symptom $S_i = s_{i,j}$ given cause $C_k = 1$ and $C_h = 0 \; \forall h \neq k$	$\sum_{i=1}^{M} R_i \cdot Q_i$
$P_{S_{i,j}\|C_0}$	$i = 1, ..., M$ $j = 1, ..., Q_i$	Probability of symptom $S_i = s_{i,j}$ given cause $C_k = 0$, $\forall C_k \in C_r^i$	$\sum_{i=1}^{M} Q_i$

Table 5.5: Example of $P(C_i|O_r^i, O_n^i)$ table (Step 1)

C	O_1 =off	O_1 =on
c_1	0.2	0.6
c_2	0.7	0.7

Model construction for SBM

SBM was depicted in Fig.5.7. In this BN the required probabilities are the prior probabilities of causes, $P(C)$, and the probabilities of symptoms and conditions given causes, $P(S_i|C)$ and $P(O_i|C)$. The data provided by experts are those in Tables 5.3 and 5.4, which should be converted into the required probabilities in the SBM.

Causes are the mutually exclusive states, $c_1, ..., c_K$, of variable C. For a cause without parent conditions the probability elicited by experts, P_{C_i}, is the probability of the cause, $P(C = c_i)$. However, when the cause depends on some conditions, experts provide the conditional probability of the cause given the related conditions. In order to define the probability table for the C node, firstly an auxiliary probability table of the cause given the conditions should be built, taken into account that $P(C_i|O_r^i, O_n^i) = P(C_i|O_r^i)$. For example, let's consider two causes C_1 and C_2, and a condition O_1, which is related to C_1. The probabilities elicited by experts are $P_{C_1|O_1=off} = 0.2$, $P_{C_1|O_1=on} = 0.6$, $P_{C_2} = 0.7$, $P(O_1 = off) = 0.3$. Then, the probability table $P(C|O_1)$ would be that shown in Table.5.5.

Taking into account that in the SBM the causes are the exclusive states of a single node, the probabilities of the causes should sum up 1. That means that the sum of probabilities of the states of the C node should be 1 for any column of the conditional probability table (any combination of states of the conditions). There are two ways of dealing with that constraint: either the expert is responsible of checking that he gave the right probabilities or it is allowed that the elicited probabilities do not comply with that constraint, and then they are modified

Table 5.6: Example of $P(C_i|O_r^i, O_n^i)$ table (Step 2)

C	O_1 =**off**	O_1 =**on**
c_1	0.15	0.46
c_2	0.54	0.54
c_3	0.31	0

Table 5.7: Probability table for S_i node in SBM

$S_i\|C$	c^i_{r1}	...	$c^i_{rR_i}$	c^i_{n1}	...	$c^i_{n(K+1-R_i)}$
$s_{i,1}$	$P_{S_{i,1\|C^i_{r1}}}$	...	$P_{S_{i,1\|C^i_{rR_i}}}$	$P_{S_{i,1\|C_0}}$	...	$P_{S_{i,1\|C_0}}$
$s_{i,2}$	$P_{S_{i,2\|C^i_{r1}}}$	...	$P_{S_{i,2\|C^i_{rR_i}}}$	$P_{S_{i,2\|C_0}}$	...	$P_{S_{i,2\|C_0}}$
...	...	...	...	...	...	...
s_{i,Q_i}	$P_{S_{i,Q_i\|C^i_{r1}}}$	...	$P_{S_{i,Q_i\|C^i_{rR_i}}}$	$P_{S_{i,Q_i\|C_0}}$	...	$P_{S_{i,Q_i\|C_0}}$

by KAT. The latter will be done according to the following procedure:

- If the sum of any column is higher than 1, the followed criterion has been to maintain the ratio amongst the probabilities of the same cause given different states of the conditions. Hence, a normalization constant B is defined as the highest of the sum of the columns. In the example in Fig.5.5, B would 1.3. Then, all entries in the probability table should be normalized by dividing them by B.

- If the sum of probabilities of any column is less than 1, a cause c_{K+1}, named "Others", is added to the table. That cause stands for any other cause of the problem not considered by the expert. The probability of that cause is 1 minus the sum of all probabilities in the same column. Fig.5.6 shows how the table 5.5 would be modified.

- The probability of each cause, $P(C = c_i)$, should be calculated according to the following equations:

$$P(C = c_i) = \frac{1}{B}\sum_{O_r^i} P_{C_i|O_r^i} \cdot \prod_{O_j \in O_r^i} P_{O_j} = \frac{1}{B} \sum_{O^i_{r_1},\ldots,O^i_{r_{U_i}}} P_{C_i|O^i_{r_1},\ldots,O^i_{r_{U_i}}} \cdot P_{O^i_{r_1}} \cdot \ldots \cdot P_{O^i_{r_{U_i}}} \quad (5.54)$$

According to the previous procedure, for the example in Table 5.5, the probabilities for the C node would be $P(C = c1) = 0.37$, $P(C = c2) = 0.54$ and $P(C = c3) = 0.09$.

Probability table for symptom S_i is presented in Table 5.7. On the one hand, the probabilities of S_i conditioned to related causes have been explicitly elicited by the expert. On the other hand, the expert has also defined the probability of the symptom conditioned to non-related causes, which is the same for all non-related causes.

In the SBM, conditions are represented as children of the parent node. Therefore, the probabilities of conditions given causes are required, whereas the available probabilities are

Table 5.8: Probability table for O_i node in CBM

O_i	$P(O_i)$
$o_{i,1}$	$P_{O_{i,1}}$
...	...
o_{i,X_i}	$P_{O_{i,X_i}}$

the prior probabilities of conditions and the probabilities of causes given conditions. Elicited probabilities can be transformed into the required ones following the Bayes' rule:

$$P(O_j = o_{j,k}|C = c_i) = \frac{P(C = c_i|O_j = o_{j,k}) \cdot P_{O_{j,k}}}{P(C = c_i)}, \; O_j \in O_r^i \tag{5.55}$$

where $P(C = c_i)$ can be calculated following eq.(5.54) and

$$P(C = c_i|O_j = o_{j,k}) = \frac{1}{B} \sum_{O_r^i \setminus O_j} P_{C_i|O_r^i} \cdot \prod_{O_h \setminus O_j \in O_r^i} P_{O_h} \tag{5.56}$$

For causes which are independent of condition O_j it is assumed that conditions are independent of each other given the causes, as suggested in [173]. Then, instead of eq. (5.55), the following equation should be used:

$$P(O_j = o_{j,k}|C = c_i) = \frac{\left(1 - \sum\limits_{C_h|O_j \in O_r^h} P(C = c_h|O_j = o_{j,k})\right) \cdot P_{O_{j,k}}}{1 - \sum\limits_{C_h|O_j \in O_r^h} P(C = c_h)}, \; O_j \in O_n^i \tag{5.57}$$

Model construction for CBM

The difference between the SBM and the CBM is how conditions are modelled. In the CBM conditions are parent of the cause node (Fig.5.8). Thus, probability tables for condition nodes are directly built with the prior probabilities of conditions elicited by experts (Table 5.8).

The probability table for the C node is defined as the auxiliary conditional probability table of the SBM (see page 137). Finally, the probability tables for the symptoms coincide with the tables in the SBM (Table 5.7).

Model construction for ICI

In models designed following the ICI assumptions, causes are modelled as different nodes (Fig.5.11). Each cause node has two states (*false*/*true*). Probability tables for condition variables are those shown in Table 5.8 for the CBM. Probability tables for the cause nodes are

Table 5.9: Probability table for C_i node in ICI

$C_i\|O_r^i$	O_r^i
false	$1 - P_{C_i\|O_r^i}$
true	$P_{C_i\|O_r^i}$

Table 5.10: Probability table for S_i node in ICI

$S_i\|C_r^i$	C_r^i
$s_{i,1}$	$1\text{-}\sum_{k=2}^{Q_i} P(S_i = s_{i,k}\|C_r^i)$
$s_{i,2}$	$\sum_{\{A\|g(A)=s_{i,2}\}} \prod_{k=r_1}^{r_{R_i}} P_{A_k\|C_k}$
...	...
s_{i,Q_i}	$\sum_{\{A\|g(A)=s_{i,Q_i}\}} \prod_{k=r_1}^{r_{R_i}} P_{A_k\|C_k}$
	$A = \{A_{r_1}, ..., A_{r_{R_i}}\}$

directly built based on the information elicited by expert (Table 5.9). Finally, probability tables for the symptom variables are defined according to eq.(5.21), as shown in Table 5.10.

If a leak cause is included in the model, the expert normally defines the probability of each related cause given that none of the other related causes are present and the leak cause is present, $P_{S_{i,j}|C_k}$. That probability can be expressed depending on the probability of the leak cause and the probabilities in eq.(5.21) as

$$P_{S_{i,j}|C_k} = \sum_{\{A_0,A_k|g(A_0,A_k,A_h|\forall h\neq k \ \ A_h=s_{h,1})=s_{i,j}\}} P_{A_0|C_0} \cdot P_{A_k|C_k} \tag{5.58}$$

From this equation, the values of $P_{A_{k,j}|C_k}$ can be obtained depending on $P_{S_{i,j}|C_k}$ and $P_{S_{i,j}|C_0} = (P_{A_{k,j}|C_0})$. For example, eq.(5.25) codes the relations among the previous probabilities for the Noisy-OR model.

5.7.2 Bayesian Classifier

Knowledge gathering

KA for the BC is composed of the phases presented in Fig.5.20. All phases but 4 and 5 are identical to the ones explained for the construction of BNs. Step 4 is also equal to the specification of probabilities for causes and conditions for BNs. However, while in BNs probabilities for symptoms had to be defined, in BCs symptoms are assumed to be continuous and the information to be provided by experts are the parameters of the pfds. As explained in 5.2.3, the conditional probabilities of the symptoms given the causes can be modelled as beta pdfs or combinations of beta pdfs. Therefore, the required data are the a and b parameters of these beta pdfs. Nevertheless, normally the experts are not familiar with the meaning of those parameters. Hence, the aim should be to identify two parameters which are easy to define by experts and could be converted into the a and b parameters.

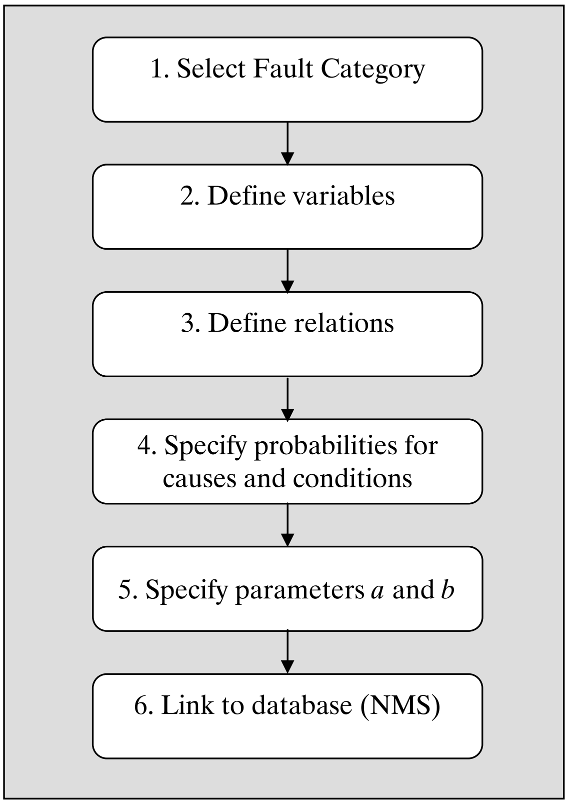


Figure 5.20: Phases in Knowledge Acquisition for BC

The first parameter could be the expected value or the mode. In some tests with experts it has been observed that for symptoms it is more intuitive to specify the mode ν, i.e. the most probable value, than the expected value.

The second parameter should give an idea of the variance of the distribution. It is proposed to request the symptom value, y, above the mode whose probability is much lower than that of the mode. For example, the question to be asked could be "what is the value of the symptom whose probability is 100 times lower than the probability of the mode?".

These two parameters should be defined for the conditional pdf of each symptom given each related cause. In addition, the parameters for the beta pdf for each symptom given that none of the related causes is present should also be specified. KAT should depict the beta pdf based on the parameters elicited by the expert, so that he/she could check whether the pdf fits with his/her knowledge.

Model construction

Probability tables for conditions and causes are built as those of the CBM. Defining pdfs for symptoms means determining the values of a and b of their corresponding beta pdfs. Different methods have been proposed for the elicitation of those two parameters by experts [198, 148]. The Equivalent Prior Sample (EPS) method asks the experts to give an estimate of the mean of the distribution, ρ, and of the sample size, n, that he is basing his assessment upon. The larger the sample size, the more information the expert believes he has. Based on his answers, the parameters of the beta distribution can be calculated as $a = n\rho$ and $b = n(1-\rho)$. In the Cumulative Distribution Function (CDF) method, the expert is asked to give his estimate of the median of the distribution and to give another quantile. The beta distribution can then be fitted to the quantiles. In the Probability Density Function (PDF) method, experts elicit the mode, ν, and a value of the symptom that is half as likely as the modal value.

The method proposed in this thesis is a generalization of the PDF method. The two parameters defined by experts are the mode, ν, and a η parameter, which is a symptom value whose probability is k times ($k < 1$) the probability of the mode[4].

The mode of the beta pdf is:

$$\nu = \frac{a-1}{a+b-2} \tag{5.59}$$

Then, the a and b parameters of the beta pdfs can be obtained from ν, k and η as follows:

$$a = \frac{log\left(\frac{\nu(1-\nu)^{\frac{1}{\nu}}(1-\eta)}{k\cdot\eta(1-\eta)^{\frac{1}{\nu}}(1-\nu)}\right)}{log\left(\frac{\nu(1-\nu)^{\frac{1}{\nu}}(1-\eta)}{\eta(1-\eta)^{\frac{1}{\nu}}(1-\nu)}\right)} \tag{5.60}$$

$$b = \frac{a-1+2\nu-\nu a}{\nu} \tag{5.61}$$

[4]The PDF method is a particularization of this method, when $k = 0.5$

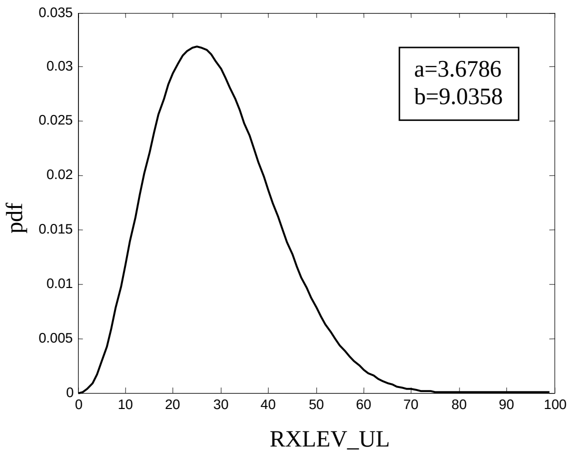


Figure 5.21: Example of definition of beta pdf

For example, in order to obtain the pdf of the symptom “RXLEV_UP” conditioned to the cause “Lack of coverage”, the expert states that the most probable value for the symptom is 25%. In addition, he/she asserts that the probability of the symptom being 70% is much lower (100 times lower) than the probability of the symptom being 25%. The resulting pdf is shown in Fig.5.21.

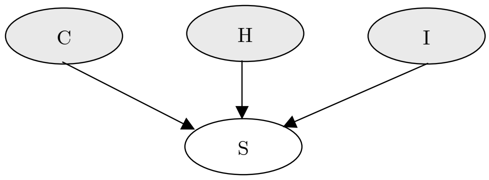


Figure 5.22: Example of ICI model

Table 5.11: Example of ICI model

$P(I)$	(30,70)
$P(C)$	(20,80)
$P(H)$	(40,60)
$P(S\|I=1,C=0,H=0)$	(10,30,60)
$P(S\|I=0,C=1,H=0)$	(20,30,50)
$P(S\|I=0,C=0,H=1)$	(5,20,75)
$P(S\|I=0,C=0,H=0)$	(100,0,0)

5.A Appendix: Example of model under ICI assumptions

A symptom S="percentage of UL quality HOs (Qua_HO_UL)" may have three possible causes: "UL interference (I)", "Lack of coverage (C)" or "Hardware fault (H)". S has been discretized in three intervals: low/medium/high. Thus, its corresponding random variable has three states: 0, 1, 2. Probabilities elicited by experts are shown in Table 5.11. First probability in the second column of the table corresponds to the distinguished state. The purpose of this example is to compare the probability table of symptom S when ICI models with different g functions are applied. The leak cause has not been included, i.e. when all causes are absent the symptom is at its distinguished state.

Fig.5.22 shows the BN structure adopted to model this example. Probability table for symptom S applying different g functions are shown in Table 5.12 (Noisy-Max), Table 5.13 (Noisy-Add), Table 5.14 (Noisy-Addlim) and Table 5.15 (Noisy-Addceil). Table 5.16 presents the probability of each cause depending on the state taken by the symptom S for all the structures.

Table 5.12: Probability table for symptom S in Noisy-Max example

I	0				1			
H	0		1		0		1	
C	0	1	0	1	0	1	0	1
0	100	20	5	1	10	2	0.5	0.1
1	0	30	20	11.5	30	18	9.5	4.9
2	0	50	75	87.5	60	80	90	95

Table 5.13: Probability table for symptom S in Noisy-Add example

I	0				1			
H	0		1		0		1	
C	0	1	0	1	0	1	0	1
0	100	20	5	1	10	2	0.5	0.1
1	0	30	20	5.5	30	9	3.5	0.85
2	0	50	75	23.5	60	26	16.5	4.6
3	0	0	0	32.5	0	33	34.5	13.6
4	0	0	0	37.5	0	30	45	27.6
5	0	0	0	0	0	0	0	30.75
6	0	0	0	0	0	0	0	22.5

Table 5.14: Probability table for symptom S in Noisy-Addlim example

I	0				1			
H	0		1		0		1	
C	0	1	0	1	0	1	0	1
0	100	20	5	1	10	2	0.5	0.1
1	0	30	20	5.5	30	9	3.5	0.85
2	0	50	75	93.5	60	89	96	99.05

Table 5.15: Probability table for symptom S in Noisy-Addceil example

I	0				1			
H	0		1		0		1	
C	0	1	0	1	0	1	0	1
0	100	20	5	1	10	2	0.5	0.1
1	0	30	20	61.5	30	68	54.5	19.05
2	0	50	75	37.5	60	30	45	80.85

Table 5.16: Probabilities (C,H,I) vs.symptom value

	Max	Add	Addlim	Addceil
0	(44.44, 6.98, 18.92)	(44.44, 6.98, 18.92)	(44.44, 6.98, 18.92)	(44.44, 6.98, 18.92)
1	(76.16, 35.94, 60.81)	(68.92, 24.13, 49.33)	(68.92, 24.13, 49.33)	(74.54, 45.41, 61.03)
2	(83.16, 67.75, 75.14)	(67.63, 39.20, 52.68)	(83.50, 67.18, 75.51)	(91.22, 84.40, 87.46)
3	-	(85.17, 62.17, 76.05)	-	-
4	-	(84.98, 73.31, 78.55)	-	-
5	-	(100, 100, 100)	-	-
6	-	(100, 100, 100)	-	-

5.B Appendix: Posterior probability of causes in SBNs

Let $S_1, \ldots, S_M$ be the discrete symptoms of a SBN. Let add M continuous children, $R_1, \ldots, R_M$, one for each discrete symptom. Then, the probability of a cause C can be calculated as:

$$P\left(C|\, R_1, R_2 \ldots R_M\right) = \frac{P(C, R_1, \ldots, R_M)}{P(R_1, \ldots, R_M)} = \frac{\sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M, R_1, \ldots, R_M)}{\sum\limits_{C} \sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M, R_1, \ldots, R_M)} = \frac{\sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M) \cdot P(R_1, \ldots, R_M | C, S_1, \ldots, S_M)}{\sum\limits_{C} \sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M) \cdot P(R_1, \ldots, R_M | C, S_1, \ldots, S_M)} \tag{5.62}$$

Taking into account that R_i is independent of all other variables when its parent S_i is known, eq.(5.62) becomes

$$P\left(C|\, R_1, R_2 \ldots R_M\right) = \frac{\sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M) \cdot P(R_1 | S_1) \ldots P(R_M | S_M)}{\sum\limits_{C} \sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M) \cdot P(R_1 | S_1) \ldots P(R_M | S_M)} = \frac{\sum\limits_{S_1 \ldots S_M} P(C, S_1, \ldots, S_M) \cdot P(R_1 | S_1) \ldots P(R_M | S_M)}{\sum\limits_{S_1 \ldots S_M} P(S_1, \ldots, S_M) \cdot P(R_1 | S_1) \ldots P(R_M | S_M)} \tag{5.63}$$

The probabilities $P(R_i|S_i)$ are the belief mapping functions f^{S_i}. Hence, if the value of the continuous symptoms $R_1, \ldots, R_M$ are $s_1, \ldots, s_M$, the posterior probability of the cause C_k is

$$P\left(C_k|\, s_1 s_2 \ldots s_M\right) = \frac{\sum\limits_{S_1 S_2 \ldots S_M} P(C_k, S_1, ..., S_M) \cdot \prod\limits_{h=1}^{M} f^{S_h}(s_h)}{\sum\limits_{S_1 S_2 \ldots S_M} P(S_1, ..., S_M) \cdot \prod\limits_{h=1}^{M} f^{S_h}(s_h)} \tag{5.64}$$

5.C Appendix: Equations for Belief mapping functions of SBNs

In the following equations a and b are the minimum and maximum value, respectively, of the continuous symptom S_i^C, T is the threshold and p is the degree of smoothness.

- **Rectangular functions (rect)**:

$$f_1^{S_i}(x) = \begin{cases} \frac{1}{T-a} & a \leq x \leq T \\ 0 & T < x \leq b \end{cases} \qquad f_2^{S_i}(x) = \begin{cases} 0 & a \leq x \leq T \\ \frac{1}{b-T} & T < x \leq b \end{cases} \tag{5.65}$$

- **Trapezoidal functions (trap)**:

$$\begin{aligned} & q = \min\left[p,\ 2\left(T-a\right),\ 2\left(b-T\right)\right] \\ & f_1^{S_i}(x) = \begin{cases} \frac{1}{T-a} & a \leq x \leq T - {}^{q}\!/_{2} \\ \frac{1}{T-a}\left(-\frac{x-T}{q}+0.5\right) & T - {}^{q}\!/_{2} < x < T + {}^{q}\!/_{2} \\ 0 & T + {}^{q}\!/_{2} \leq x \leq b \end{cases} \\ & f_2^{S_i}(x) = \begin{cases} 0 & a \leq x \leq T - {}^{q}\!/_{2} \\ \frac{1}{b-T}\left(\frac{x-T}{q}+0.5\right) & T - {}^{q}\!/_{2} \leq x \leq T + {}^{q}\!/_{2} \\ \frac{1}{b-T} & T + {}^{q}\!/_{2} \leq x \leq b \end{cases} \end{aligned} \tag{5.66}$$

- **Rect-gaussian functions (gaus)**:

$$\begin{aligned} & q = \min\left[p,\ 2\left(T-a\right),\ 2\left(b-T\right)\right] \\ & f_1^{S_i}(x) = \begin{cases} \frac{1}{k_1} & a \leq x \leq T - {}^{q}\!/_{2} \\ \frac{1}{k_1}\exp\left(-\frac{(s-T+0.5q)^2}{2q^2}\cdot 4Ln\ 4\right) & T - {}^{q}\!/_{2} \leq x \leq b \end{cases} \\ & k_1 = T - {}^{q}\!/_{2} - a + {}^{q\sqrt{2\pi}}\!/_{4\sqrt{Ln\ 4}} - {}^{q\sqrt{2\pi}}\!/_{2\sqrt{Ln\ 4}} \cdot Q\left(\frac{b-T+{}^{q}\!/_{2}}{{}^{q}\!/_{2}\sqrt{Ln\ 4}}\right) \\ & f_2^{S_i}(x) = \begin{cases} \frac{1}{k_2}\exp\left(-\frac{(s-T-0.5q)^2}{2q^2}\cdot 4Ln\ 4\right) & a \leq x \leq T + {}^{q}\!/_{2} \\ \frac{1}{k_2} & T + {}^{q}\!/_{2} \leq x \leq b \end{cases} \\ & k_2 = b - T - 0.5q - {}^{q\sqrt{2\pi}}\!/_{4\sqrt{Ln\ 4}} + {}^{q\sqrt{2\pi}}\!/_{2\sqrt{Ln\ 4}} \cdot Q\left(\frac{a-T-{}^{q}\!/_{2}}{{}^{q}\!/_{2}\sqrt{Ln\ 4}}\right) \end{aligned} \tag{5.67}$$

Table 5.17: Example of qualitative model

Type	Variable	States
Fault Category	F_1 =High DCR	
Causes	C_1 =Interference C_2 =Lack of coverage C_3 =Hardware fault	false/true false/true false/true
Symptoms	S_1 =Q1_7_UL S_2 =Q1_7_DL S_3 =RXLEV_UL S_4 =RXLEV_DL S_5 =Link_imb	normal/high normal/high normal/high normal/high normal/high
Conditions	O_1 =FH O_2 =Cell type	off/on sparse/normal/dense
Causes related to symptoms	$C_r^1 = \{C_1, C_2, C_3\}$ $C_r^2 = \{C_1, C_2, C_3\}$ $C_r^3 = \{C_2, C_3\}$ $C_r^4 = \{C_2, C_3\}$ $C_r^5 = \{C_3\}$	
Conditions related to causes	$O_r^1 = \{O_1\}$ $O_r^2 = \{O_2\}$	

5.D Appendix: Example of KA

An expert in troubleshooting of GERAN has specified the qualitative model (Table 5.17) and the quantitative model (Table 5.18). From this information, a diagnosis model is built. Fig.5.23 shows a BN following the SBM structure. Probability tables are shown in Tables 5.19, 5.20 and 5.21. Cause $C = c_4$ represents any other cause of the problem not explicitly defined by the expert.

Fig.5.24 shows an alternative model, which follows the CBM structure. Probability tables for symptom nodes are the same ones as for the SBM (Table 5.20). Tables 5.22 and 5.23 show the probability table for the cause node and the condition nodes, respectively.

Fig.5.25 depicts a Noisy-OR model. Probability tables for condition nodes are those in Table 5.23. Probability tables for the cause and the symptom nodes are presented in Tables 5.24 and 5.25, respectively.

The results provided by these three models were compared. Table 5.26 presents the probabilities of the causes obtained with each model under different evidence. The cause with the highest probability (in bold in the table) was selected as the diagnosed cause. It can be observed that when the cause is clear, probabilities are similar for all BNs (first example on the table). However, when it is not clear that only a cause is present (second example on the table), results are different for models under ICI assumptions. This is due to the fact that both the SBM and the CBM assumes that only a cause is present at a time, whereas more than a cause may be present when using the Noisy-OR model.

Table 5.18: Example of quantitative model

Parameters	Value
Thresholds	$t_{1,1} = 12\%$ $t_{2,1} = 19\%$ $t_{3,1} = 16\%$ $t_{4,1} = 10\%$ $t_{5,1} = 23\%$
Cause Prob.	$P_{C_1\|O_1=o_{1,1}} = 0.3$ $P_{C_1\|O_1=o_{1,2}} = 0.2$ $P_{C_2\|O_2=o_{2,1}} = 0.5$ $P_{C_2\|O_2=o_{2,2}} = 0.4$ $P_{C_2\|O_2=o_{2,3}} = 0.05$ $P_{C_3} = 0.2$
Cond. Prob.	$P_{O_{1,i}} = (0.3, 0.7),\ i = \{1, 2\}$ $P_{O_{2,i}} = (0.4, 0.3, 0.3),\ i = \{1, 2, 3\}$
Symptom Prob.	$P_{S_{1,2\|C_1}} = 0.7$ $P_{S_{1,2\|C_2}} = 0.3$ $P_{S_{1,2\|C_3}} = 0.5$ $P_{S_{2,2\|C_1}} = 0.7$ $P_{S_{2,2\|C_2}} = 0.2$ $P_{S_{2,2\|C_3}} = 0.5$ $P_{S_{3,2\|C_2}} = 0.5$ $P_{S_{3,2\|C_3}} = 0.5$ $P_{S_{4,2\|C_2}} = 0.7$ $P_{S_{4,2\|C_3}} = 0.5$ $P_{S_{5,2\|C_3}} = 0.5$ $P_{S_{1,2\|C_0}} = 0.1$ $P_{S_{2,2\|C_0}} = 0.1$ $P_{S_{3,2\|C_0}} = 0.1$ $P_{S_{4,2\|C_0}} = 0.1$ $P_{S_{5,2\|C_0}} = 0.1$

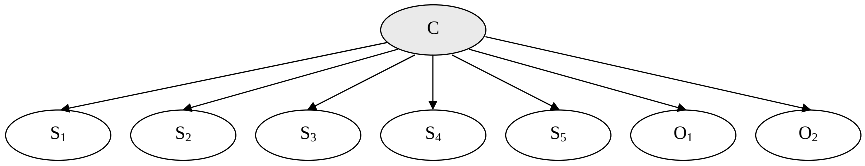


Figure 5.23: Example of SBM structure

Table 5.19: Probability table for cause node in SBM

C	**P(C)**
c_1	0.23
c_2	0.335
c_3	0.2
c_4	0.235

Table 5.20: Probability table for symptom nodes in SBM

$S_1\|C$	c_1	c_2	c_3	c_4
$s_{1,1}$	0.3	0.7	0.5	0.9
$s_{1,2}$	0.7	0.3	0.5	0.1
$S_2\|C$	c_1	c_2	c_3	c_4
$s_{2,1}$	0.3	0.8	0.5	0.9
$s_{2,2}$	0.7	0.2	0.5	0.1
$S_3\|C$	c_1	c_2	c_3	c_4
$s_{3,1}$	0.9	0.5	0.5	0.9
$s_{3,2}$	0.1	0.5	0.5	0.1
$S_4\|C$	c_1	c_2	c_3	c_4
$s_{4,1}$	0.9	0.3	0.5	0.9
$s_{4,2}$	0.1	0.7	0.5	0.1
$S_5\|C$	c_1	c_2	c_3	c_4
$s_{5,1}$	0.9	0.9	0.5	0.9
$s_{5,2}$	0.1	0.1	0.5	0.1

Table 5.21: Probability table for condition nodes in SBM

$O_1\|C$	c_1	c_2	c_3	c_4
$o_{1,1}$	0.3913	0.2727	0.2727	0.2727
$o_{1,2}$	0.6087	0.7273	0.7273	0.7273
$O_2\|C$	c_1	c_2	c_3	c_4
$o_{2,1}$	0.3007	0.5970	0.3007	0.3007
$o_{2,2}$	0.2707	0.3582	0.2707	0.2707
$o_{2,3}$	0.4286	0.0448	0.4286	0.4286

Table 5.22: Probability table for cause node in CBM

O_1	$o_{1,1}$		$o_{1,2}$		$o_{1,3}$	
O_2	$o_{2,1}$	$o_{2,2}$	$o_{2,1}$	$o_{2,2}$	$o_{2,1}$	$o_{2,2}$
c_1	0.3	0.2	0.3	0.2	0.3	0.2
c_2	0.5	0.5	0.4	0.4	0.05	0.05
c_3	0.2	0.2	0.2	0.2	0.2	0.2
c_4	0	0.1	0.1	0.2	0.45	0.55

Table 5.23: Probability table for condition nodes in CBM

O_1	$P(O_1)$
$o_{1,1}$	0.3
$o_{1,2}$	0.7

O_2	$P(O_2)$
$o_{2,1}$	0.4
$o_{2,2}$	0.3
$o_{2,3}$	0.3

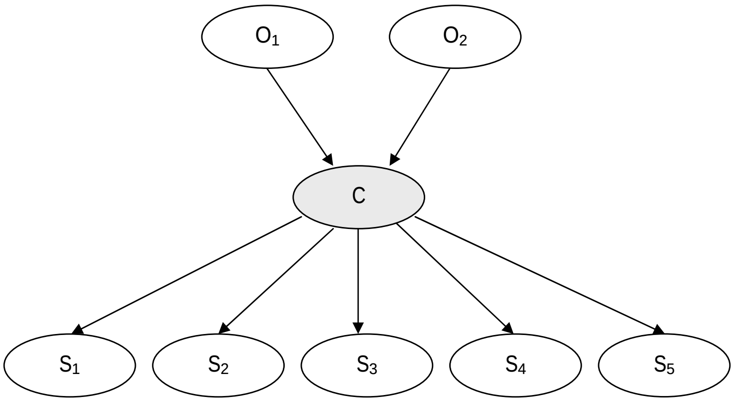


Figure 5.24: Example of CBM structure

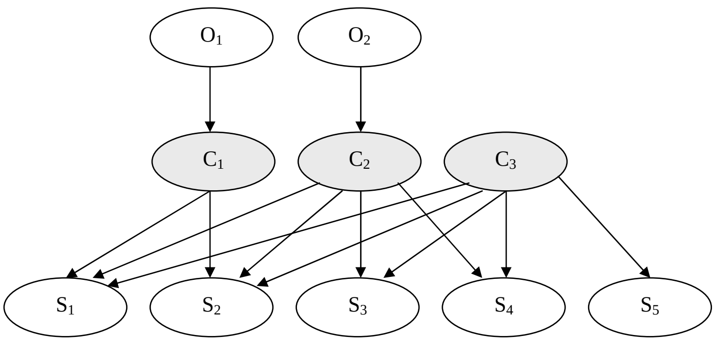


Figure 5.25: Example of Noisy-OR structure

Table 5.24: Probability table for cause nodes in Noisy-OR model

$C_1 \mid O_1$	$o_{1,1}$	$o_{1,2}$
off	0.7	0.8
on	0.3	0.2

$C_2 \mid O_2$	$o_{2,1}$	$o_{2,2}$	$o_{2,3}$
off	0.5	0.6	0.95
on	0.5	0.4	0.05

C_3	
off	0.8
on	0.2

Table 5.25: Probability table for symptom nodes in Noisy-OR

S_1								
C_3	off				on			
C_2	off		on		off		on	
C_1	off	on	off	on	off	on	off	on
$s_{1,1}$	0.9	0.3	0.7	0.23	0.5	0.17	0.39	0.13
$s_{1,2}$	0.1	0.7	0.3	0.77	0.5	0.83	0.61	0.87

S_2								
C_3	off				on			
C_2	off		on		off		on	
C_1	off	on	off	on	off	on	off	on
$s_{2,1}$	0.9	0.3	0.8	0.2667	0.5	0.1667	0.4444	0.1481
$s_{2,2}$	0.1	0.7	0.2	0.7333	0.5	0.8333	0.5556	0.8519

S_3				
C_3	off		on	
C_2	off	on	off	on
$s_{3,1}$	0.9	0.5	0.5	0.2778
$s_{3,2}$	0.1	0.5	0.5	0.7222

S_4				
C_3	off		on	
C_2	off	on	off	on
$s_{4,1}$	0.9	0.3	0.5	0.1667
$s_{4,2}$	0.1	0.7	0.5	0.8333

S_5		
C_3	off	on
$s_{5,1}$	0.9	0.5
$s_{5,2}$	0.1	0.5

Table 5.26: Probability of causes

Evidence			
$S_1 = s_{1,2}$	$S_2 = s_{2,2}$	$S_3 = s_{3,1}$	$S_4 = s_{4,1}$
$S_5 = s_{5,1}$	$O_1 = o_{1,2}$	$O_2 = o_{2,3}$	
Cause	**SBM**	**CBM**	**Noisy-OR**
C_1	**0.8929**	**0.8701**	**0.8597**
C_2	0.0037	0.0049	0.0143
C_3	0.0812	0.0761	0.1200
C_4	0.0222	0.0488	-

Evidence			
$S_1 = s_{1,2}$	$S_2 = s_{2,2}$	$S_3 = s_{3,2}$	$S_4 = s_{4,2}$
$S_5 = s_{5,1}$	$O_1 = o_{1,2}$	$O_2 = o_{2,1}$	
Cause	**SBM**	**CBM**	**Noisy-OR**
C_1	0.0431	0.0531	0.5538
C_2	**0.6384**	**0.5696**	**0.8770**
C_3	0.3174	0.3767	0.4492
C_4	0.00107	0.0005	-

Part III

Evaluation

Chapter 6

Results

This chapter presents the evaluation of the methods and models proposed in this thesis. In the first part of the chapter, Section 6.1, the experimental design is described. In Section 6.1.1, it is explained how cases used to test and train the different models have been obtained. Two types of data have been used: cases from a live GERAN network and simulated cases. The figures of merits defined to evaluate and compare the diverse systems are presented in Section 6.1.2. In addition, the sensitivity to imprecision in thresholds and probabilities was analyzed for each model. Section 6.1.3 explains how sensitivity analysis was performed. Section 6.1.4 describes the methodology adopted for the evaluation of the different systems.

The second part of the chapter summarizes the main achieved results. In Section 6.2, it is described how the parameters for a model from which simulated cases could be generated were obtained. In Section 6.3, the experience with mobile network operators and manufacturers in knowledge acquisition and in testing of a troubleshooting prototype is briefly presented. In Section 6.4, diverse diagnosis models were evaluated: Bayesian Classifier, BNs with different structures, SBNs and MUI. In addition, the methods proposed in Chapter 5 to learn the parameters of the model (thresholds and probabilities) were compared. Finally, the drawbacks and benefits of all models are discussed in Section 6.5.

6.1 Experimental design

6.1.1 Network and simulated cases

One commonly accepted criterion for testing new classifiers is the use of real data from the application domain to validate the systems. With this purpose, some repositories of machine learning databases have been created and maintained [50]. Unfortunately, there are no such databases of classified cases for fault diagnosis in cellular systems. In a cellular network, the NMS keeps historical records of the value of performance indicators and configuration parameters for all cells under its supervision. However, in case there is a problem in a cell, the diagnosed cause is not saved in the NMS once the fault is solved. Although most operators keep track of the faults by means of a *trouble ticket* system (see Chapter 2), this feature is not integrated in

Table 6.1: Cases from a live GSM/GPRS network

Cause	**N^{er} of cases**
UL interference	123
DL interference	122
Lack of coverage	188
HW fault	105
Transmission fault	70
Transcoder fault	35
Total	**643**

the NMS.

Therefore, one of the main difficulties in this thesis, as well as in the creation of an application to be used by operators, has been to evaluate the diagnosis systems. Ideally, the systems should be tested in a live network. This implies that a network operator should grant access to the NMSs and/or other databases of performance indicators and configuration parameters. In addition, everyday troubleshooting experts should assert which are the cells with problems and, once the fault is solved, they should save the cause together with its symptoms and conditions. On the one hand, comparison with the cause diagnosed by the automatic system would allow to measure the performance of the system under test and to fine-tune the parameters in the model. Furthermore, a database of diagnosed cases would be built, composed of cases with the cause, conditions and symptoms for each problematic cell. This database could be used to evaluate new classifiers. On the other hand, learning of model parameters (thresholds and probabilities) requires large sets of classified cases, which should be also obtained from a live network. In spite of these benefits, operators are reluctant to give access to their databases because they do not want information about network performance to be leaked out. On the one hand, without inputs from the operator the troubleshooting tool will not work well. On the other hand, operators do not want to invest any time as the tool is not working well.

Due to the fact that real cases are essential for design and evaluation of diagnosis systems, a three month trial was carried out in a live GSM/GPRS network. The network was composed of about 25000 cells and 10 NMSs. Everyday some problematic cells were identified and their faults were manually diagnosed by experts in troubleshooting. The average values of the main performance indicators on the day the fault occurred, together with the diagnosed cause, were saved in a database of classified cases. In this way, more than 600 labelled cases were obtained. The number of cases per fault cause is presented in Table 6.1.

These cases suffer from some limitations. Firstly, there is a lack of information: neither alarms nor configuration parameters were saved, not all symptoms in the model in Chapter 4 were considered, cases were classified into only a few causes, e.g. the different HW faults were not distinguished. Secondly, the cause diagnosed by the experts was not validated, therefore some cases might be erroneously diagnosed. Finally, cells were not randomly selected, therefore the percentage of cases per cause does not necessary correspond to the prior probabilities of the causes.

Table 6.2: Symptoms in the diagnosis model

Number	Code	Number	Code
Dropped Calls			
2	RF_dcr	3	RF_HO_dcr
4	A-bis_dcr	8	Tr_dcr
10	LAPD_dcr		
Level and Quality			
11	Q1_7_UL	12	Q1_7_DL
13	Q0_UL-Q0_DL	14	Q0_DL-Q0_UL
15	RXLEV_UL	16	RXLEV_DL
17	Link_imb	18	Q567_Lev_UL
19	Q567_Lev_DL		
Timing Advance			
21	TA_10		
Interference			
24	IOI		
Handovers			
26	Lev_HO_UL	27	Lev_HO_DL
28	Qua_HO_UL	29	Qua_HO_DL
30	Int_HO_UL	31	Int_HO_DL
33	In_HO		
RACH			
34	RACH		

Due to these constraints, a simplified diagnosis model has been designed. Alarms and conditions were not included in the model. In addition to causes presented in Table 6.1 another cause named “Others” has been included. Symptoms are shown in Table 6.2. The labels (number and code) of the symptoms in Table 4.10 have been maintained.

In addition, in order to palliate the limitations listed above, the following assumptions were considered:

- The spatial and temporal statistical characterization of the network is the same. In other words, the cases can be obtained not only from a given cell along different days, but also from different cells in the network. The reason for that supposition was the need to get as many cases as possible in a short time.

- It was supposed that the diagnosis asserted by troubleshooting experts was correct, although it was not checked.

The number of cases was still considered scarce compared to other domains in which BN have been used. Thus, an algorithm to simulate as many cases as required was believed to be essential. Taking into account the previous two assumptions, the procedure to generate simulated cases was as follows:

1. Statistical estimation:

- According to 5.2.3, the conditional pdfs of continuous symptoms given the causes may be accurately approximated by beta pdfs. The parameters of the beta pdfs were obtained using a maximum likelihood estimate of the classified cases from the real GSM network.
- Prior probabilities of causes were elicited by diagnosis experts.

2. Case generator:

- For each case, the cause was generated according to the prior probabilities of causes.
- Once the cause was set, the value of each symptom was obtained from the previous defined conditional pdfs by using rejection or inversion methods [112].

In this algorithm, two assumptions are implicit: i) only one cause is present at a time; ii) symptoms are independent given the cause. The first assumption is usually true in the domain under study, whereas the last one is not valid for most symptoms and causes. Nevertheless, numerous studies have proven that even if strong attribute dependencies exist, models which follow those assumptions perform very well [74, 165, 82]. Thus, the impact of ii) is expected to be insignificant.

Following this method, two types of sets were generated: *training sets*, which were used by the different learning algorithms to estimate the parameters in the models; and *test sets*, which were used to measure the performance of the diagnosis systems.

6.1.2 Performance measures

The probabilistic diagnosis systems presented in this thesis provide a probability distribution over the set of possible causes, conditioned to a set of conditions and symptoms. The classification decision is based on this probability distribution. For example, the decision could be selecting the cause with the highest probability. Another criterion could be specifying a threshold and decide that the cause is present if its probability is higher than the threshold. The former has been the decision criterion taken in the evaluation of diagnosis systems in this thesis.

Let D be a test set of N cases. In machine learning literature, the problem of assessing the performance of a classifier is based on a validation process in which the classifier is applied to the test set D and the output provided by the classifier for each of the cases in the test is compared with the actual class. Based on that comparison, the following figures of merit can be defined:

- The **diagnosis error** is the percentage of cases in D which are incorrectly diagnosed. Complementarily, the **diagnosis accuracy** is the percentage of cases correctly classified.
- For each case, the **correct cause probability** is the probability provided by the diagnosis system for the real cause. The average and standard deviation of this parameter over the N cases is calculated. This measure is very important because it may indicate the level of confidence in the diagnosis. For example, if the probability of the most probable cause

was 0.3, the diagnosis would be insecure. In addition, if the decision criterion is based on a probability threshold, the classifier would be better if the correct cause probability was high. Furthermore, most algorithms for decision-theoretic troubleshooting require the calculus of the probabilities of the causes [96, 174].

- For each case, the **correct diagnosis probability** is the average of the probability asserted by the diagnosis system for the true cause and one minus the probability provided for each of the other causes. The average and standard deviation of this parameter over the N cases is calculated. This figure takes into account that, ideally, not only the probability of the correct cause should be high, but also the probabilities asserted for the other causes should be low.

- For each case, the **rank order** is the order number of the real cause in the list of causes sorted according to their probabilities. The average and standard deviation of this parameter over the N cases is calculated. The **rank figure** provides an indication of the rank order as a percentage of the first element in the ordered list of probabilities. That is, the rank figure is 100 when the diagnosed cause coincides with the real cause and it is zero when the probability of the diagnosed cause is the lowest in the list. The rank figure can be calculated as:

$$F = 100 \cdot \frac{K - r}{K - 1} \tag{6.1}$$

where K is the number of causes and r is the rank order.

Furthermore, other figures were defined in order to evaluate the accuracy separately for the diverse causes. Let D_k^r be a subset of D composed of the cases whose real cause is C_k and let D_k^n be a subset of D composed of the cases whose real cause is not C_k, $D_k^n = D \backslash D_k^r$. Then, the following figures have been measured:

- The **type I error** is the percentage of false negatives, that is the percentage of cases in D_k^r incorrectly classified.

- The **type II error** is the percentage of false positives, that is the percentage of cases in D_k^n whose diagnosed cause is C_k.

- When the decision criterion is to assert that a cause is present when its probability is over a given threshold P, all previous figures of merits can still be used. Furthermore, a measure sometimes found in literature is the **area under the ROC[1] curve** [178, 91]. The ROC curve plots the percentage of true positives (=1-type I error) as a function of the percentage of false positives (=type II error) as the threshold P is varied from 1 to 0. Each point of the curve characterizes the performance of the classifier for a given threshold. The area under the ROC curve can be used as a measure of classification performance. An area

[1]ROC curve: Receiver operating characteristic curve

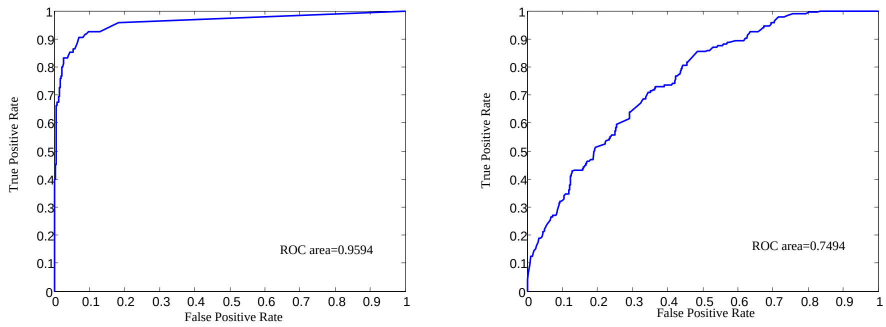


Figure 6.1: Examples of ROC curves

of 1 corresponds to perfect accuracy. An area of 0.5 corresponds to the performance of a classifier that randomly classifies the cases. Fig.6.1 shows two examples of ROC curves and their corresponding areas. The classifier whose ROC curve is on the left provides better results than that whose ROC curve is depicted on the right.

6.1.3 Sensitivity Analysis

The figures of merit in Section 6.1.2 were used to evaluate the performance of the diagnosis systems proposed in this thesis. In addition, an important aspect to consider is the sensitivity of the results to changes in the parameters (probabilities and thresholds) of the models. It is common that the parameters in the model are not those which would achieve the best classification accuracy, either because of the availability of a scarce number of training cases in data-based models or because of imprecision in the elicited parameters in knowledge-based systems. This is the reason why it is so important to analyze how changes in the specified parameters would impact the diagnosis results.

Basically, there are two approaches to sensitivity analysis: theoretical and empirical. The theoretical approach expresses the posterior probability of interest in terms of the parameters under study. The empirical methods examine the effects of varying the parameters of the network on the diagnosis. In the latter case, the most frequent approach consists of adding random noise to the probabilities and examining the effects on the diagnostic performance [103, 118, 158].

The sensitivity of BNs to imprecise probabilities have been extensively analyzed in the literature [55, 126, 158], however no references can be found about the sensitivity of the results of BNs when the discretization uses inaccurate thresholds. The sensitivity analysis to imprecise thresholds presented in this thesis is based on previous empirical methods for the sensitivity analysis of BNs to probabilities. That is, random noise at increasing levels is added to the parameters and the effects on the performance is evaluated. In the diagnosis model for cellular networks, most symptoms are measured as the percentage of samples complying with a certain condition (see 5.2.3). Therefore, the thresholds, like the probabilities, are in the [0,100] range. In other case, the threshold could be scaled to be in the [0,100] range, the noise would then be

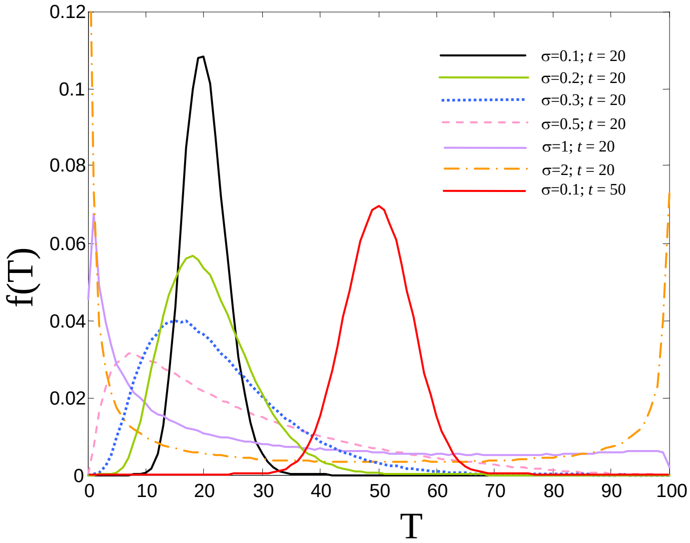


Figure 6.2: Log-odd normal distribution centered around the nominal threshold 20 or 50

added and the threshold would be re-scaled to its original range.

Adding random noise directly to the parameters has two problems. Firstly, a large additive error may produce a parameter greater than 100 or lower than 0. Secondly, the same error value seems more serious in parameters near 0 or 100 than in those in the middle range, because the resulting parameter could be very close to one of the limits (leading to a probability equal to zero or one in the case of probabilities and to a single state of the discretized symptom in the case of thresholds). In order to overcome those problems, it has been proposed to add noise to the log-odd rather than directly to the probabilities [158] or thresholds. The new parameter with added noise, T, in terms of the nominal parameter t and noise ε is

$$T = \frac{100}{1 + \left(\frac{100}{t} - 1\right) \cdot 10^{-\varepsilon}} \quad ; \quad \varepsilon = \text{Normal}(0,\sigma) \tag{6.2}$$

where parameters are expressed in % and ε follows a normal distribution.

Figure 6.2 shows the log-odd normal distribution centered around the nominal parameter of 20% and 50% for various values of the standard deviation σ of the noise. [118] argues that this distribution is only adequate for standard deviations lower than one. Effectively, it can be observed that for values of $\sigma > 1$ the distribution becomes bimodal, which is equivalent to consider an expert who assesses a threshold or probability near 0 and erring in judgment by such margin that the best parameter may be close to 100.

6.1.4 Methodology

This section explains the methodology followed in the evaluation of the methods presented in Chapters 4 and 5.

Firstly, a diagnosis model based on the knowledge of GSM troubleshooting experts was built with the help of an automatic knowledge acquisition tool. Tables A.1-A.3 in Appendix A present the parameters of the knowledge-based model built in this thesis. The values in the tables will be those referred as "experts" in the following sections. Table A.1 shows the prior probabilities of causes. Table A.2 shows the conditional probabilities of symptoms (which were modelled as 2-state variables) given causes. The probabilities in the table are the probabilities of the symptoms being in the state $s_{i,2}$ given that the corresponding cause, C_i, is present and all other related causes are absent. The thresholds for the symptoms are presented in Table A.3.

In order to evaluate the different diagnosis systems, it was required either to have a live network and operators available every time new systems had to be tested or to have a large database of classified cases. The former was not possible, therefore the procedure explained on page 159 was followed to obtain as many simulated cases as necessary. The 643 cases from the live GSM/GPRS network (Table 6.1) were the basis used by the algorithm. Section 6.2 presents the estimated pdfs which were used to generate the simulated cases and it evaluates the reasonability of those pdfs.

For the evaluation of the diagnosis systems, two types of test sets were defined: network cases and simulated cases. Let's denote *network cases* (net) to the 643 diagnosed cases from the live GSM/GPRS network (Table 6.1). Let's denote *simulated cases* (sim) to the cases generated following the algorithm explained on page 159.

Regarding the parameters of the models, two types of parameters were used: reference parameters and learnt parameters. For systems based on BNs, let's name *reference parameters* (ref) to the probabilities and thresholds learnt from a single training set of 5000 cases using the m-estimate and the SEMD methods, to calculate probabilities and thresholds, respectively. This size, 5000, was chosen because it is large enough so that the diagnosis error was not due to the size of the training set. For systems based on the BC, the *reference pdf* are the beta pdf obtained using a maximum likelihood estimate of the network cases. The a and b parameters of those beta pdfs (Tables A.4 and A.5) are named *reference parameters* (ref) of the BC. Let's denote *learnt parameters* (lrn) to the parameters learnt from a training set following any of the algorithms proposed in Chapter 5.

The methodology was the following (Fig.6.3):

- Selection of the type of model parameters, either reference parameters or learnt parameters. In the latter case, n training sets of size $Ntrain$ are independently generated following the algorithm on page 159. Subsequently, n sets of model parameters are calculated based on each of the n training sets. The learning algorithm may be any of the ones proposed in this thesis.

- Selection of the type of test set, either network cases or simulated cases. In the latter case, n test sets of size $Ntest$ cases are independently generated following algorithm on page 159.

- Each of the n test sets[2] is used together with one of the n sets of parameters, in order to

[2] If the test set is the one from the network, then the same test set is used for the n diagnosis.

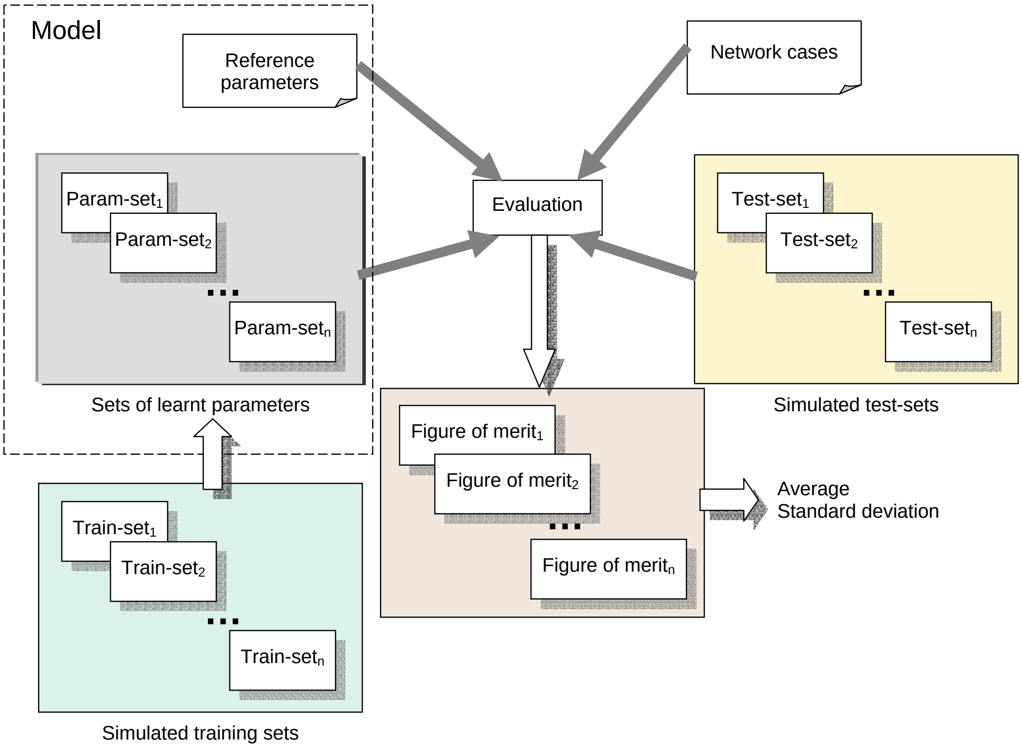


Figure 6.3: Experimental methodology

obtain a set of n diagnosis. Then, the figures of merit (average and standard deviation) are calculated for each of the n results.

Some of the evaluated aspects of the diagnosis systems proposed in this thesis have been:

1. **Figures of Merit**:

   - Figures of merits were obtained when using the system to diagnose the network cases. The parameters for the model were the reference parameters.
   - Figures of merits were obtained when using the system to diagnose simulated cases. The size of the test set was $Ntest = 1000$ cases. The parameters of the model were calculated based on a training set of $Ntrain$ cases ($Ntrain = 50$ or 5000). These experiments were repeated n times with different training sets and test sets and the average figures of merit and standard deviations over the n results were computed.

2. **Sensitivity Analysis:**

   - The sensitivity of the results to imprecise probabilities/thresholds was studied. The parameters of the model were the reference parameters. A simulated test set (i.e. $n = 1$) of size $Ntest = 1000$ cases was used. Independent log-odd random noise at increasing level was added to the probabilities/thresholds. For each level of noise, the experiment was repeated $Ntimes$ times with independently generated noise. The average figures of merit and standard deviations over the $Ntimes$ experiments were calculated for each level of noise. The study was carried out separately for the prior probabilities of causes, the conditional probabilities of symptoms given causes, the thresholds and all probabilities and thresholds at the same time.

3. **Training set size:**

   - The influence of the size of the training set on the performance was analyzed. The parameters of the model were calculated based on training sets of increasing number of cases, from $Ntrain = 100$ to 5000. The figures of merits were obtained when using the system to diagnose a simulated test set of size 1000 cases. For each size of the training set, the experiments were repeated $n = 10$ times with different test sets and training sets. The average figures of merit and standard deviations were calculated over the 10 experiments.

4. **Parameter learning:**

   - Discretization methods were compared using training sets of size $Ntrain = 50$ or 5000 cases. Tests were carried out over network cases and simulated cases (of size $Ntest = 1000$ cases). These experiments were repeated n times with different training sets and test sets (for *net* cases, the test set was always the same one) and the average figures of merit and standard deviations over the n results were computed.
   - Methods to calculate the probabilities in BNs were compared using training sets of size $Ntrain = 50$ or 5000 cases. Tests were carried out over network cases and simulated cases (of size $Ntest = 1000$ cases). These experiments were repeated n times with different training sets and test sets (for *net* cases, the test set was always the same one) and the average figures of merit and standard deviations over the n results were computed.

Furthermore, particular aspects of each diagnosis system were measured. Table 6.3 summarizes the designed diagnosis systems and the types of evaluation tests carried out for each of them. The diagnosis system under test is written on the first column. The first line of each entry in the table is the type of test set, *Cases*: either from the live network (net) or simulated cases (sim). *Param* refers to how the model parameters were obtained: either reference parameters (ref) or learnt parameters (lrn). The learning algorithm, when applicable, is denoted as *Alg*. When all methods for learning of probability and threshold were evaluated, it is denoted as *all*. The number of cases in the training set, $Ntrain$, and in the test set, $Ntest$, are other data in the table. The number of times the experiments were repeated is denoted as n. In addition, in the column "Sensitivity Analysis", it is also specified how many times the experiments were repeated for each level of noise (*Ntimes*) and the type of analysis carried out.

Table 6.3: Diagnosis systems and evaluation tests

<table>
<tr><th></th><th>Figures
of Merit</th><th>Sensitivity
Analysis</th><th>Training
set size</th><th>Discretiz.
method</th><th>Probability
Definition</th></tr>
<tr><td rowspan="2">BC</td><td>Cases=net
Param=ref</td><td rowspan="2">Cases=sim
Param=ref
Ntest=1000

n=1
Ntimes=100
Analysis:
mean /
std /
mean+std</td><td rowspan="2">Cases=sim
Param=lrn
Ntest=1000
Ntrain=100-5000
n=10
Alg=MLE</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>Cases=sim
Param=lrn
Ntest=1000
Ntrain=50/5000
n=500
Alg=MLE</td></tr>
<tr><td rowspan="2">SBM</td><td>Cases=net
Param=ref</td><td rowspan="2">Cases=sim
Param=ref
Ntest=1000

n=1
Ntimes=50
Analysis:
thresholds /
prior prob. /
cond prob. /
thres.+prob.</td><td rowspan="2">Cases=sim
Param=lrn
Ntest=1000
Ntrain=100-5000
n=10
Alg=all</td><td>Cases=net
Param=lrn

Ntrain=50/5000
n=500
Alg=all</td><td>Cases=net
Param=lrn

Ntrain=50/5000
n=500
Alg=all</td></tr>
<tr><td>Cases=sim
Param=lrn
Ntest=1000
Ntrain=50/5000
n=50
Alg=SEMD+mest</td><td>Cases=sim
Param=lrn
Ntest=1000
Ntrain=50/5000
n=500
Alg=all</td><td>Cases=sim
Param=lrn
Ntest=1000
Ntrain=50/5000
n=500
Alg=all</td></tr>
<tr><td rowspan="2">N-OR</td><td>Cases=net
Param=ref</td><td rowspan="2">Cases=sim
Param=ref
Ntest=1000

n=1
Ntimes=50
Analysis:
thresholds /
prior prob. /
cond prob. /
thres.+prob.</td><td rowspan="2">Cases=sim
Param=lrn
Ntest=1000
Ntrain=100-5000
n=10
Alg=SEMD+mest</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>Cases=sim
Param=lrn
Ntest=1000
Ntrain=50/5000
n=50
Alg=SEMD+mest</td></tr>
<tr><td rowspan="2">SBN/
MUI</td><td>Cases=net
Param=ref</td><td rowspan="2">Cases=sim
Param=ref
Ntest=1000
n=1
Ntimes=50
Analysis:
thresholds /
prior prob. /
cond prob. /
thres.+prob.</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td>Cases=sim
Param=lrn
Ntest=1000
Ntrain=50/5000
n=50
Alg=SEMD+mest/exp</td></tr>
</table>

6.2 Estimation of probability functions

In order to obtain as many cases as necessary, the procedure explained on page 159 was followed. The first step approximated the conditional pdfs of symptoms given causes by beta pdfs. With that purpose, the cases from the trial in a live GSM/GPRS network (Table 6.1) were classified depending on their causes, C_k, in subsets D_k. The procedure to calculate the parameters a and b of the beta pdfs was as follows:

- For each subset D_k, repeat:
  - For each symptom S_i, calculate $a(i,k)$ and $b(i,k)$ of the beta pdf by using a maximum likelihood estimate [164] of the values of the symptom, $\{s_i^{(1)}, ..., s_i^{(M_k)}\}$, where M_k is the number of cases in subset D_k.

For a given cause C_k and a symptom S_i not related to C_k, the percentage of cases where the symptom value is equal to zero is normally high. This is the reason why, instead of using just a beta pdf, the pdf of the symptom was approximated by a sum of a delta at zero and a beta pdf. If the value of the delta is close to one, the remaining number of cases (i.e. those cases where the symptom is different to zero) will be small and, in many cases, they will not be enough to approximate the pdf by a beta pdf. The estimated parameters, which are denoted as *reference parameters*, are shown in Tables A.4 and A.5. These values were used to build the reference pdfs, which were taken as the gold standard to generate the simulated cases. The parameters in the table are (a,b,d), where the pdf is defined as

$$f_{S_i|C_k}(x) = d \cdot \delta(x) + (1-d) \cdot f_\beta(x,a,b) \tag{6.3}$$

where $f_\beta(x,a,b)$ is a beta pdf, with parameters a and b, in x.

Probability plots (P-P plots) were used in order to assess whether the estimated beta pdfs were reasonable models for the data. Probability plotting is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data [141, 130]. To construct a P-P plot, the observations in the sample are first ranked from smallest to largest, $x_{(1)}, x_{(2)}, ..., x_{(N)}$. The hypothesized cumulative distribution function (cdf) at each point is then plotted against their observed cdf $i/N, i = 1, ..., N$. If the hypothesized distribution adequately describes the data, then the plotted points will fall approximately along a straight line; if the plotted points deviate significantly from a straight line, then the hypothesized model is not appropriate.

In addition, chi-square tests [141, 130] were applied to evaluate the goodness of fit[3]. The k intervals for the chi-square tests were chosen so that the number of points in each bin was higher

[3]**Chi-square tests:** For each estimated pdf, the N samples were arranged in a histogram of k bins. Let O_i be the observed frequency in the ith interval. From the hypothesized probability distribution, e.g. the beta pdf, the expected frequency in the ith bin, denoted E_i, is computed. E_i can be calculated from the probabilities of each interval given by the hypothesized pdf, p_i, as Np_i. The test statistic is:

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(O_i - Np_i)^2}{Np_i} \tag{6.4}$$

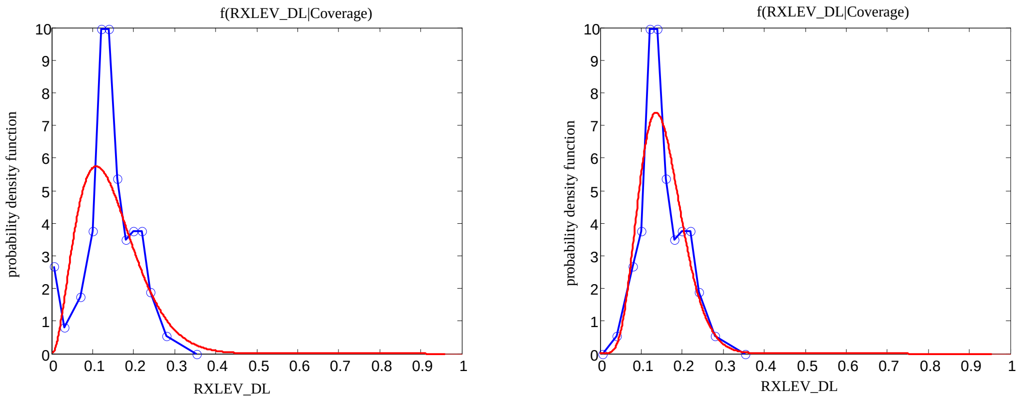
or equal to 5. In addition, the maximum number of intervals was set to 50. The significance level was set to 1%. The results of the chi-square tests were positive for most cause / symptom pairs, although there were some exceptions. The estimated beta pdf for the symptom 16, "RXLEV_DL", given cause "Lack of coverage" did not pass the chi-square test. This is due to the fact that there were four samples (2.1% of the samples) with too low value, as can be appreciated in Fig.6.4(a). In the figure, the normalized histogram for the data from the cellular network is depicted in blue and the estimated pdf is drawn in red. There might be several reasons for those low samples: the total number of samples was very low, different types of cells were not distinguished, etc. If the samples lower than 3% were considered to be zeros, the beta pdf that would be obtained would pass the chi-square test. In that case, the histogram would be that in Fig.6.4(b). The parameters obtained assuming those samples to be zero were the ones adopted to generate the simulated cases. Likewise, the pdf of the same symptom, "RXLEV_DL", given causes "hardware" and "transmission fault" and the pdf of symptom "RXLEV_UL" given causes "hardware", "lack of coverage" and "transmission fault", although passed the chi-square test, could be better fitted if low value samples were not considered in the approximation. Fig.6.5-6.9 shows the histogram built without rejecting low value samples (on the left) and assuming that samples lower than 3% are zero (on the right). Table A.6 shows the parameters of the beta pdfs for both options.

Symptom 17, "Link_imb", can be modelled as described on page 104. The histograms and the modelled pdfs for that symptom given the different causes are shown in Fig.6.10. Instead of the previous pdfs, in order to simplify the model, the pdfs of symptom 17 conditioned to causes have been modelled as beta pdfs (in black in Fig.6.10). P-P plots and chi-square tests have demonstrated that the approximation with beta pdfs is quite good. For some of the causes it can be observed that the approximations with beta pdfs is even more accurate than those following page 104.

Figures in Appendix B shows in blue the normalized histograms for the data from the cellular network and in red the estimated pdfs for all pairs symptom / cause. The bins for the histograms were in each case those used for the chi-square tests, that is, the number of samples in each bin was at least 5. Corresponding P-P plots (for the beta component of the pdf) are also depicted on the figures. Delta components are not displayed on the figure. Symptoms whose delta was higher than 0.9 are not represented in Appendix B.

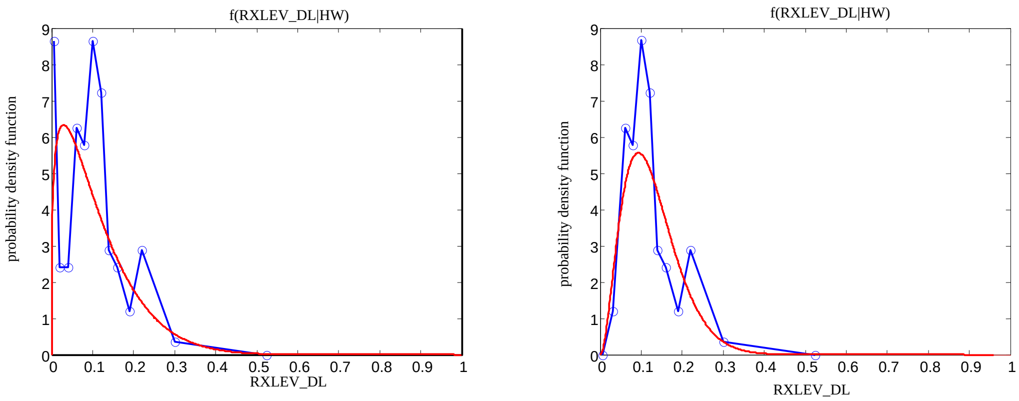
Finally, pdfs with parameters defined in Tables A.4 and A.5, which were used to generate the simulated cases, are presented in Appendix C. Those figures can be used to determine which causes are related to each symptom. For example, Fig.6.11 shows the percentage of samples in the UL with level lower than -100 dBm given the diverse causes. In the figure, it can be noticed that the value of this symptom is increased when there is a lack of coverage or a hardware fault.

If the population follows the hypothesized distribution, X_0^2 follows, approximately, a chi-square distribution with $k-p-1$ degrees of freedom, where p represents the number of parameters of the hypothesized distribution estimated by sample statistics, which is 2 in the case under study (a and b). The hypothesis that the distribution of the population is the hypothesized distribution should be rejected if the calculated value of the test statistic $X_0^2 > X^2_{\alpha,k-p-1}$, where α is the significance level, that is the probability of rejecting the hypothesis when it is true.



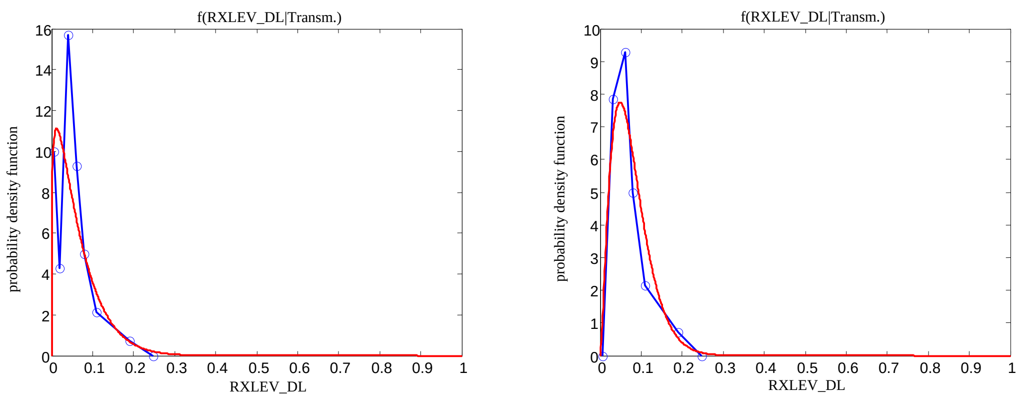
(a) (b)

Figure 6.4: Symptom “RXLEV_DL” given “Lack of coverage”



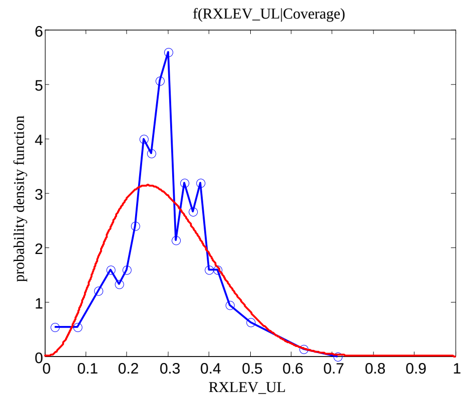
(a) (b)

Figure 6.5: Symptom “RXLEV_DL” given “Hardware fault”

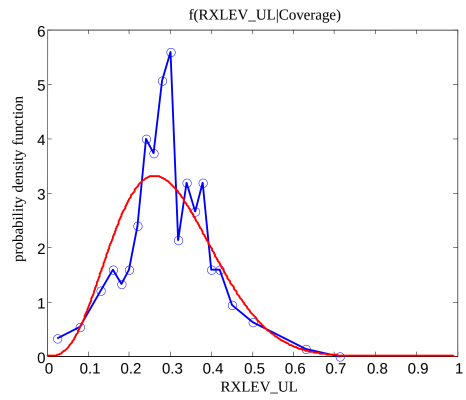


(a) (b)

Figure 6.6: Symptom “RXLEV_DL” given “Transmission fault”

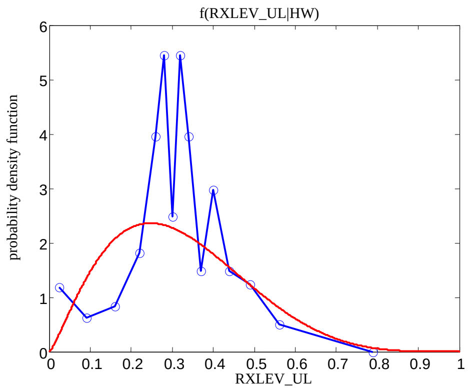


(a)

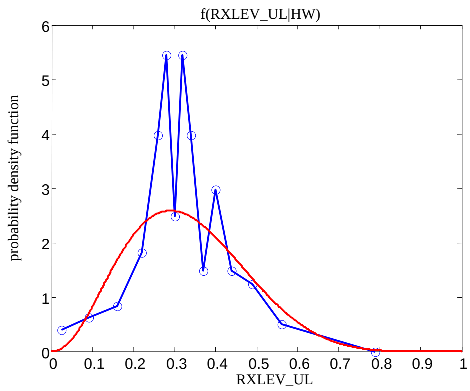


(b)

Figure 6.7: Symptom “RXLEV_UL” given “Lack of coverage”

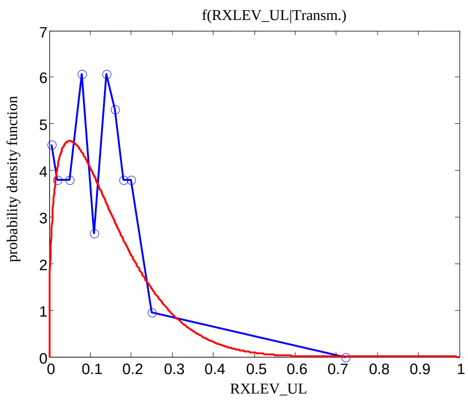


(a)

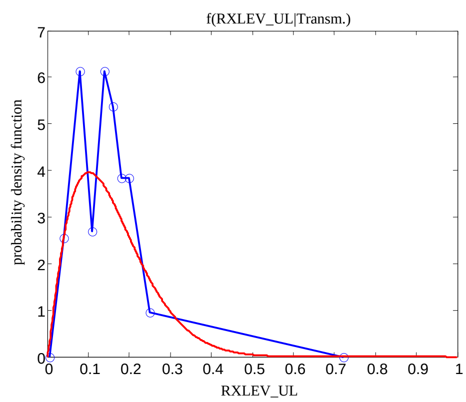


(b)

Figure 6.8: Symptom “RXLEV_UL” given “Hardware fault”



(a)



(b)

Figure 6.9: Symptom “RXLEV_UL” given “Transmission fault”

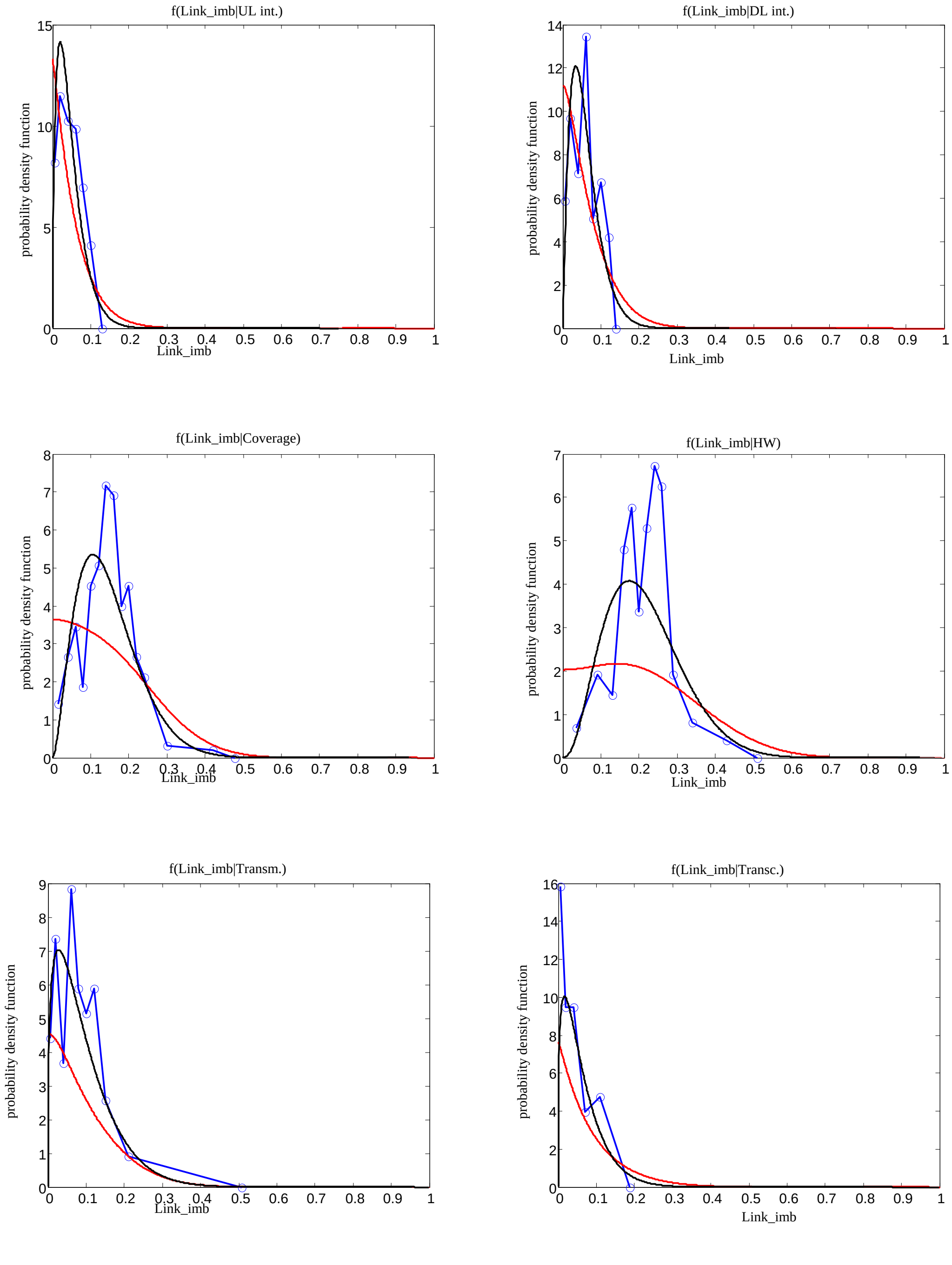


Figure 6.10: Symptom “Link_imb”

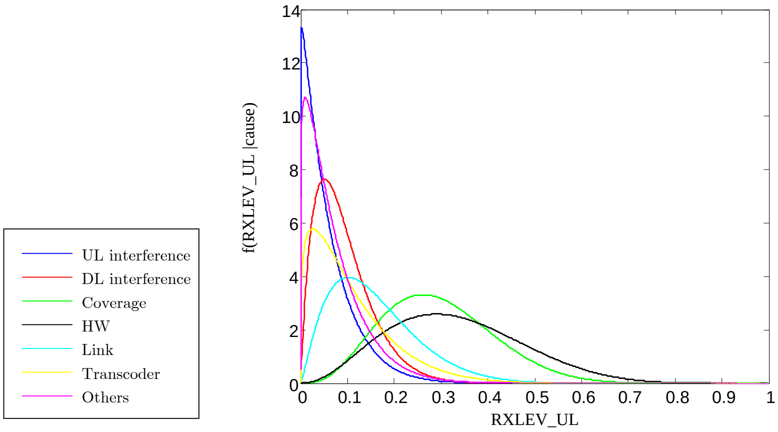


Figure 6.11: Symptom RXLEV_UL given causes

6.3 Prototype tools and field tests

During years 2001 and 2002, a cooperation agreement was set between a network operator (who's name has to be kept confidential), a manufacturer of network equipment (Nokia) and an university department (Ingeniería de Comunicaciones, University of Málaga) in order to develop and test some prototype tools and models for diagnosis of cellular networks. In this project, my responsibility was the research part of the project and the specifications of the prototype tools. In addition, I was also involved in other tasks, such as follow-up of trials and management meetings. The main results of that project will be briefly summarized in this section. They comprises the implementation of two prototype tools: a knowledge acquisition tool and a troubleshooting tool, and the evaluation of the tools and models in a live network. It should be pointed out that by that time this thesis was in its beginnings and most of the methods proposed in it were still not developed. Thus, the deployed system was based on knowledge only.

6.3.1 Knowledge acquisition

A knowledge acquisition tool, which was named KAT [36, 87], was built following the method explained in Section 5.7. The aim of KAT was to assist network planners and troubleshooting experts in the creation of diagnosis models. The user could define models without the requirement of prior knowledge on BNs thanks to an user interface. From the information supplied, KAT automatically created BNs with two possible structures: SBM and Noisy-OR.

One of the key issues for testing KAT with the experts was the usability. That is, the user interface of the first versions of KAT was so complex that the experts preferred to use a spreadsheet for model specification. Therefore, much of the effort was applied to simplifying the user interface, taking into account the users' opinion. Fig.6.12 shows one snapshot of KAT corresponding to the "Project Window".

Another problem was found in the definition of probabilities for conditions. At the beginning,

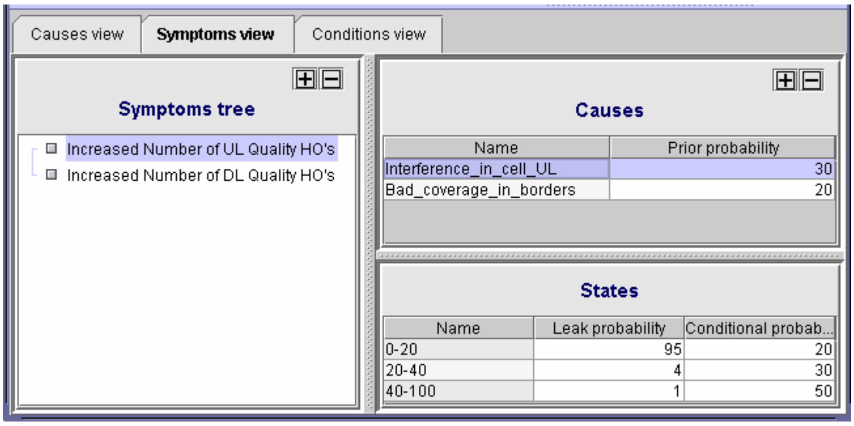


Figure 6.12: KAT: Project window with symptom view selected

the user was required to specify prior probabilities of conditions, prior probabilities of causes and conditional probabilities of causes given conditions. This definition is redundant because probabilities of causes can be obtained from probabilities of conditions and probabilities of causes given conditions. The reason for asking for the three types of probabilities was to homogenize probability definition, that is, in this way prior probabilities of causes were specified for all causes, not only for those causes without parent conditions. After the user specified the probabilities, conditional probabilities were automatically changed so that the probabilities of causes were the defined ones by minimizing the square root error of the defined conditional probability [37]. In the tests with the users, this methodology was perceived as very confusing because they did not understand why the defined probabilities were changed. Therefore, the conclusion was that, even if the definition was not homogenous for all causes, it was better not to ask for prior probabilities of causes that had parent conditions.

6.3.2 Troubleshooting Tool

Within the Nokia agreement, a prototype troubleshooting tool (TST) [39, 108] was also developed (Fig.6.13). The user had to supply the following information to the TST:

- Fault case, e.g. congestion, high drop call-rate, etc., which determined the Bayesian model to use.
- Cells to be analyzed (e.g. top-10 bad performing cells according to statistics collected). Normally this is automatically fed by the Fault Detection system.
- Date of the analysis, together with the averaging method used to calculate the performance statistics (busy hour, 24 hours data, ...).

When the previous information was entered into the tool, the user was asked for those data that could not be automatically collected from network statistics (e.g. weather condition, cell

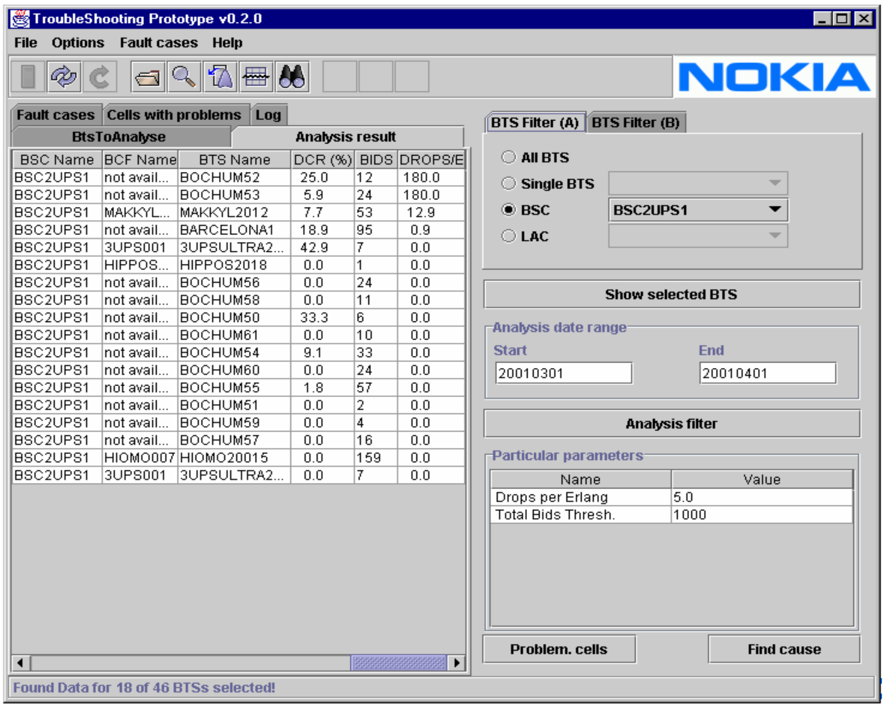


Figure 6.13: Snapshot of prototype TST

density, etc.). Next step was to run the analysis or schedule it, e.g. run TST during early hours of the morning so that analysis was ready when engineers came in the morning. During the analysis the TST collected data from all sources specified in the model (databases, alarms, etc.). When the data collection finished, the TST performed the reasoning, based on the selected Bayesian model, and it calculated the probability of each possible cause being the one causing the problem(s). The TST presented, for each cell, the list of possible causes of the problem ranked by their efficiency (function of the probability and the cost of the action required to solve the problem).

All the collected evidence related to symptoms could be displayed or saved locally in text format. The user could also save his own feedback on the problem cause together with the results of the analysis. This information could be utilized to verify whether the solution provided by the tool was the right one.

6.3.3 Trial in a live GSM/GPRS network

As part of the Nokia agreement, a model for diagnosis in GSM/GPRS was developed based on the expertise of troubleshooters from the operator, on knowledge of personnel from network planning services in Nokia and on the guidance from people in the project at Nokia and at the university. The TST was deployed in a live network and the model was fine tuned while it was tested to diagnose faulty cells. The causes, symptoms and conditions included in the model are presented in Tables D.1, D.3 and D.2, respectively.

The trial[4] in the live network had a duration of about six months. Results of the trial are presented in two phases: "Trial 1", which includes partial results at the middle of the trial period; and "Trial 2", which includes final results at the end of the trial period. The main evaluated aspects were the diagnosis accuracy, the time to carry out troubleshooting and the usability of the tools.

Troubleshooting experts analyzed the faulty cells and the assessed diagnosis was compared with that provided by the TST. Cases were classified in four groups:

- Unknown diagnosis
  - On going: Analysis was started by the experts, but it was not completed, i.e. the correct cause was not yet known.
  - Inconclusive: The real cause of the problem was unclear and it was not possible to sort it out by the experts. Therefore, it was not possible to know if the analysis done by the TST was correct.
- Known diagnosis
  - Correct: The analysis done by the TST gave the right cause, which was verified and confirmed by the experts.
  - Incorrect: The analysis done by the TST provided the wrong cause, when the experts knew the correct answer.

Figure 6.14 shows all causes proposed by the TST, classified into the previous four categories. Results show that there were 23 cases out of 55 where the correct cause was known by the experts. The right cause was proposed 11 times (47.8%) by the TST and in 12 times the TST gave incorrect causes (52.2%). A deeper analysis is presented in Table 6.4 for those 23 cases where the user knew the correct answer. The column *Number of cases* shows the number of cells in which the corresponding cause was identified. The label for the cause is indicated by a number, whose corresponding cause is shown in Table D.1. From those results, it can be observed that for some causes (12, 24 and 27) the diagnosis accuracy was 100%. On the contrary, for some causes (9 and 25) the performance of the TST was very bad, indicating some problems in the model.

Consequently, the model was fine-tuned during the following three months. The results at the end of that period are shown in Fig.6.15 and Table 6.5. From these results, it can be observed that the amount of cases where the cause was known by the expert was increased from 42% to 66%. It can also be noticed that the number of diagnosed causes in those cases (9) was higher than that of first period (7). As in results on the first period, there were some causes which were very well spotted by the TST. However, almost no improvement was achieved for causes 9 and 25. The total diagnosis accuracy was increased from 48% to 61%.

[4] This trial is different to the one presented on page 158

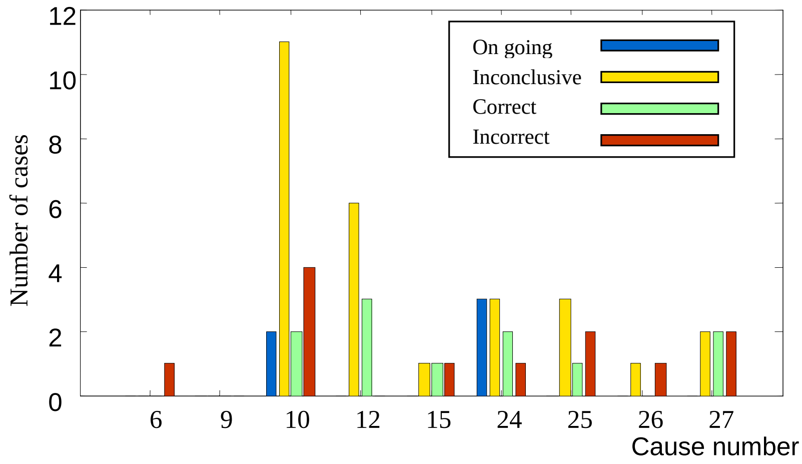


Figure 6.14: Trial results for first period

Table 6.4: Trial results for first period

Correct cause	**Ner cases**	**Cause proposed by TST** 6	9	10	12	15	24	25	26	27	**Diagn. accur.**
9	3		0					2	1		0%
10	3			2			1				67%
12	3				3						100%
15	2	1				1					50%
24	2						2				100%
25	8			4		1		1		2	13%
27	2									2	100%
Total	23	1	0	6	3	2	3	3	1	4	48%

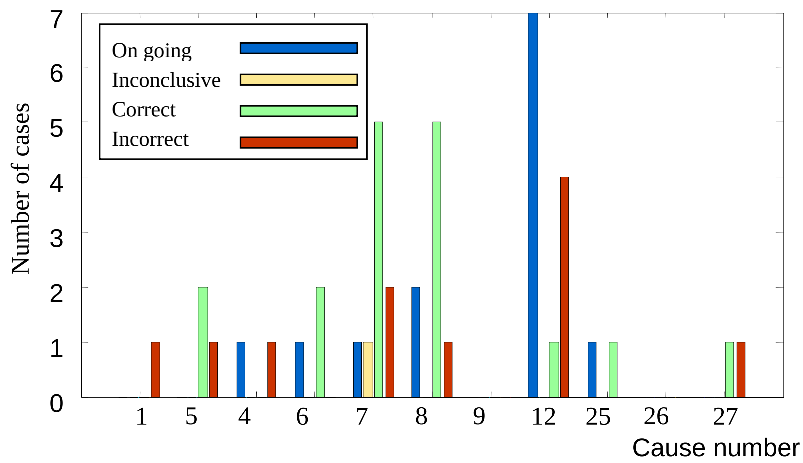


Figure 6.15: Trial results for second period

Table 6.5: Trial results for second period

Correct	**Ner**	**Cause proposed by TST**											**Diagn.**
cause	**cases**	1	5	4	6	7	8	9	12	25	26	27	**accur.**
5	4		2			1			1				50%
6	2				2								100%
7	5					5							100%
8	6						5		1				83%
9	4			1		1	1	0	1				0%
12	1								1				100%
25	4		1						1	1		1	25%
26	1	1									0		0%
27	1											1	100%
Total	28	1	3	1	2	7	6	0	5	1	0	2	61%

The achieved diagnosis accuracy (61%) was considered to be quite promising by all partners (operator, manufacturer and researchers). It was realized that fine-tuning of the model was essential, not only to adjust the parameters in the model, but also to identify new causes and symptoms. However, the fine-tuning process was considered as very time-consuming.

Another analyzed aspect of the TST performance was the time it took to make the analysis. During the trial it was measured that the time that the TST took to carry out the diagnosis was from one minute to five minutes per cell, depending on the NMS load. Compared to the regular human troubleshooting time, this significantly reduced on regular human troubleshooting time, which can last several days for the most complex cases. Even with an incorrect cause suggestion by the TST, the analysis results provide extremely valuable clues to the troubleshooting experts. This proves to be quite large time saving.

Regarding the usability of the tool, the general feeling was that the TST was easy to use and no major problems were encountered. The feedback provided on the usability was used to highly improve the TST.

Finally, the trial allowed to get a basic idea of the fault causes that could be investigated further. The TST presented a summary of the abnormal symptoms, which helped to verify the diagnosed cause. Therefore, the automatic diagnosis system was recognized as very valuable and time-saving by all partners.

6.4 Evaluation of the diagnosis systems

6.4.1 Bayesian Classifier

Figures of merit

The performance of the BC was studied for two different types of test sets: network cases and simulated cases. Table 6.6 shows the description of the experiments.

Table 6.7 shows the obtained results for Experiment 1. The diagnosis accuracy, 71.2%, is considered very good and slightly higher than that achieved by a human expert. Furthermore,

Table 6.6: BC: Description of experiments 1 and 2

Experiment 1	
Model parameters	reference parameters
Training set	Not applicable
Learning algorithm	Not applicable
Test set	network cases
Number of sets (n)	1
Experiment 2	
Model parameters	learnt parameters
Training set	Ntrain=50/5000 cases
Learning algorithm	MLE
Test set	sim: Ntest=1000 cases
Number of sets (n)	500

Table 6.7: BC: Figures of merit for network cases (mean, std)

	Figure of merit
Accuracy	71.23
Correct cause prob. (%)	(70.32, 40.15)
Correct diagn.prob. (%)	(91.41, 12.00)
Rank order	(1.5505, 1.2120)

the time required by the automated diagnosis system (seconds) is much lower than that needed by a human expert (one hour in average, even days in some cases) to provide a diagnosis. The results per cause are depicted in Fig.6.16. In that figure, it can be observed (type I error) that most errors in classification occur when the fault cause is a HW fault or a transmission fault. This is reasonable, since HW and transmission related alarms have not been considered by the designed diagnosis system. If those data had been available, the whole performance of the diagnosis system would have been considerably improved. From Fig.6.16, it can also be appreciated (type II error) that most incorrectly classified cases are diagnosed as having a lack of coverage.

Results for Experiment 2 are presented in Table 6.8. Comparing the obtained accuracy when varying the size of the training set (Ntrain= 50 or 5000), it can be concluded that the system performance is highly degraded when the number of training examples is scarce. This pinpoints a drawback of diagnosis systems based on the BC, which will be further investigated in the following section: although the achieved accuracy could be high, to do this a large database of training cases is required. Average figures of merit per fault cause are shown in Fig.6.17. As expected, for small training sets, type I errors are especially high (almost 100%) for those causes whose prior probabilities are lower.

Dependency with the size of the training set

The following experiment (Table 6.9) studied the dependency of the performance on the size of the training set. With that purpose, training sets of increasing sizes were generated, from 100 to

Figure 6.16: BC: Figures of merit per cause for network cases

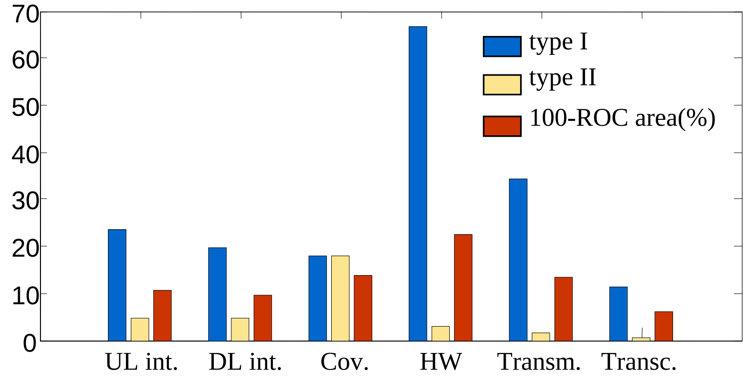


Table 6.8: BC: Average figures of merit for simulated cases (mean, std)

	|Train.set|=50	|Train.set|=5000
Accuracy	42.52	88.58
Correct cause prob. (%)	(34.87, 46.06)	(85.60, 25.98)
Correct diagn.prob. (%)	(51.89, 44.14)	(95.87, 7.51)
Rank order	(1.7578, 2.525)	(1.1408, 0.4724)

Figure 6.17: BC: Average figures of merit per cause for simulated cases

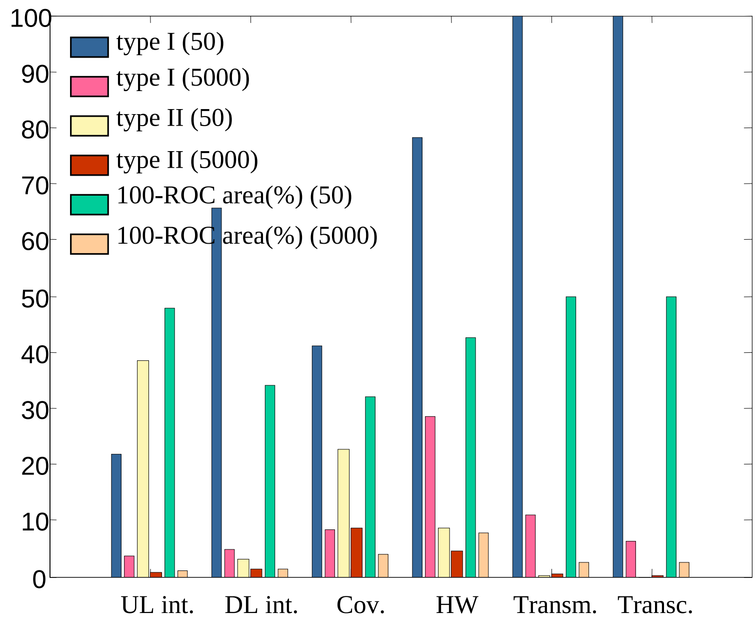
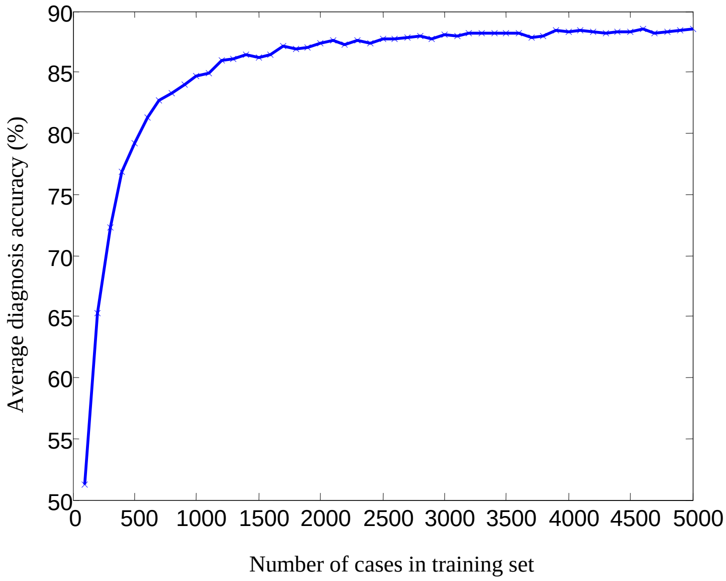


Table 6.9: BC: Description of Experiment 3

Experiment 3	
Model parameters	learnt parameters
Training set	From Ntrain= 100 to 5000 cases
Learning algorithm	MLE
Test set	sim: Ntest= 1000 cases
Number of sets (n)	10

Figure 6.18: BC: Diagnosis accuracy vs.size of training set



5000 cases. Results are presented in Fig.6.18. The strong dependency of the performance with the size of the training set can be clearly appreciated in the figure. In this experiment, from 500 cases downwards the accuracy rapidly decreases.

Sensitivity analysis

The sensitivity of the BC to imprecise definition of the model parameters was also empirically measured (Table 6.10). Log-odd normal random noise was added to the mean and/or to the standard deviation of the beta pdfs, emulating the effect of imprecision in their definition (see Section 6.1.3). Fig.6.19 shows the average diagnosis accuracy depending on the level of noise. It can be observed that, for large levels of noise, the accuracy is more sensitive to imprecision in the mean of the beta pdfs that to imprecision in the standard deviation. However, for low level of noise, the sensitivity is very similar. It can be concluded that the BC is quite sensitive to imprecision in the parameters of the model. The accuracy for no noise is 82%. For small imprecision ($\sigma = 0.1$) in both the mean and the standard deviation, accuracy is 70.5%. For $\sigma = 0.2$, the accuracy is 50% and for $\sigma = 0.3$, it is already 33%.

Table 6.10: BC: Description of Experiment 4

Experiment 4	
Model parameters	reference parameters
Training set	Not applicable
Learning algorithm	Not applicable
Test set	sim: Ntest= 1000 cases
Number of sets (n)	1
Ntimes	100

Figure 6.19: BC: Sensitivity in diagnosis accuracy

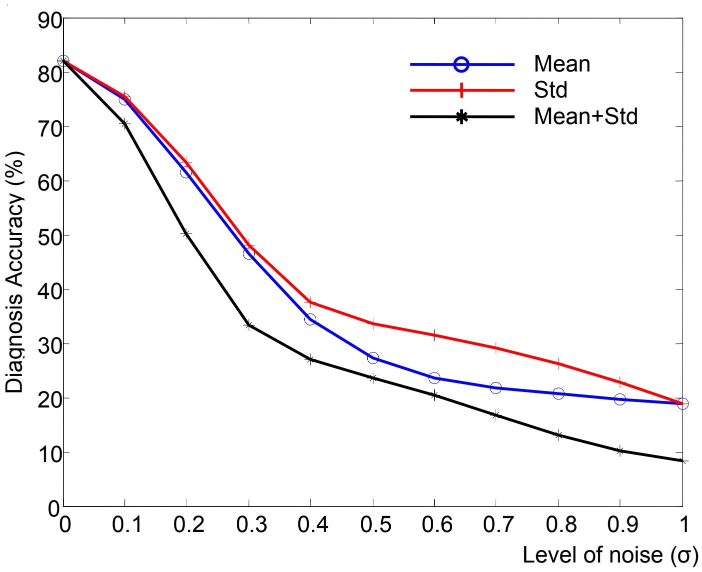


Table 6.11: BN structures: Description of experiments 1 and 2

Experiment 1	
Model parameters	reference parameters
Training set	Not applicable
Learning algorithm	Not applicable
Test set	network cases
Number of sets (n)	1
Experiment 2	
Model parameters	learnt parameters
Training set	Ntrain=50/5000 cases
Learning algorithm	SEMD and m-est
Test set	sim: Ntest=1000 cases
Number of sets (n)	50

Summary

When the parameters of the BC are accurate, either because they have been obtained based on a large database of training examples or they have been precisely elicited by experts, the diagnosis accuracy obtained with the BC is very high. The main drawback of the BC is that performance rapidly decreases if the parameters are incorrect due the the inherent uncertainty of diagnosis experts or due to a reduced number of training examples.

6.4.2 SBM versus Noisy-OR

Figures of merit

The aim of the experiments (Table 6.11) presented in this section was to compare the performance of two diagnosis system: the first one was based on a BN with a SBM structure and the second one was based on a BN following Noisy-OR assumptions [45]. For both models, symptoms were modelled as binary random variables.

Results of Experiment 1 are shown in Table 6.12, whereas figures per cause are depicted in Fig.6.20. Diagnosis accuracy is only slightly higher for the Noisy-OR (68.3%) than for the SBM (67.2%). Thus, the conclusion is that the behavior is similar for both structures[5].

Results for Experiment 2 are presented in Table 6.13. Again, the achieved accuracy for both BN structures is so close that it is not possible to state that one structure is better than the other one.

Therefore, the conclusion is that the SBM is preferred to the Noisy-OR because results are similar and the former is much simpler. However, in the test sets, only a single cause was taken place at the same time. In the future it should be tested whether the Noisy-OR outperforms the SBM in the simultaneous presence of more than a fault.

[5]In order to test the statistical significance of the results, a test to compare two proportions [141] was applied. The null hypothesis was that the Noisy-OR model achieved the same performance as the SBM. The value of the test statistic was 0.41, which is within the interval, (1.96,-1.96), calculated for a confidence level of 95%. Therefore, according to the test, we cannot reject the null hypothesis. That means that there is insufficient evidence to support the claim that the Noisy-OR model outperforms the SBM.

Table 6.12: BN structures: Figures of merit for network cases (mean, std)

	Reference	
	SBM	**Noisy-OR**
Accuracy	67.19	68.27
Cor.cause prob. (%)	(65.04, 38.65)	(68.79, 36.51)
Cor.diagn.prob. (%)	(90.01, 11.04)	(88.48, 10.17)
Rank order	(1.61, 1.17)	(1.67, 1.30)

Figure 6.20: BN structures: Figures of merit per cause for network cases

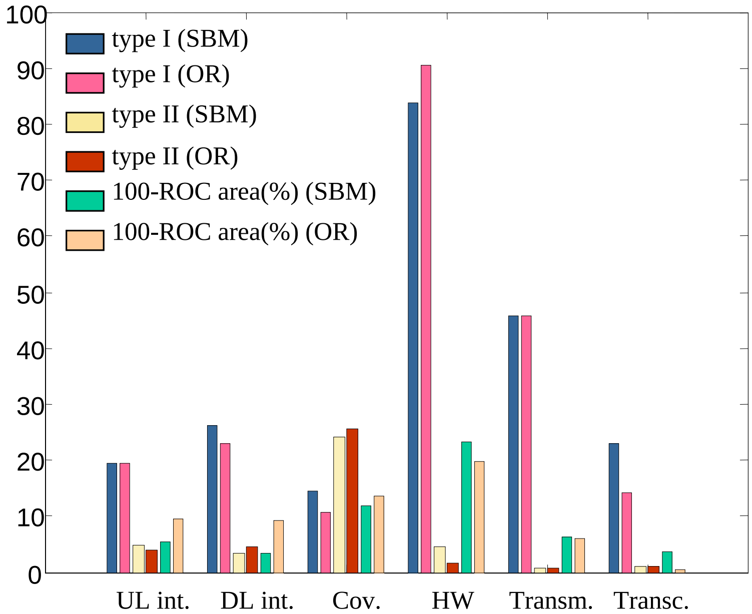


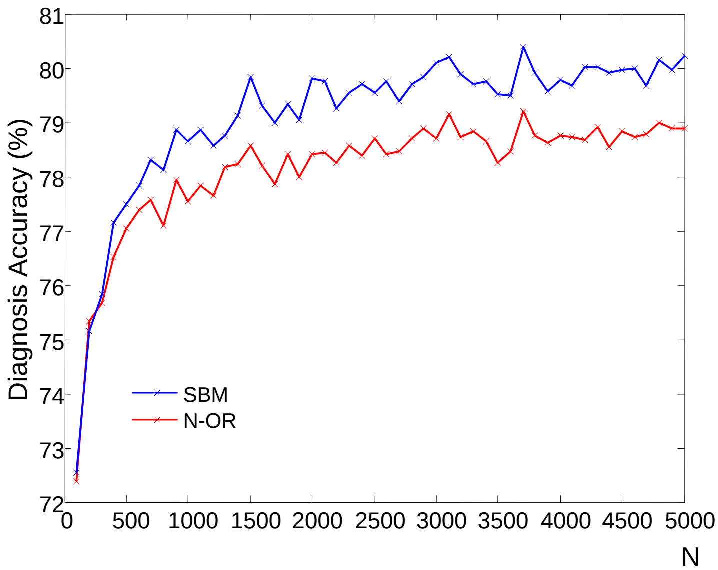
Table 6.13: BN structures: Average figures of merit for simulated cases

	|Train.set|=50		|Train.set|=5000	
	SBM	**Noisy-OR**	**SBM**	**Noisy-OR**
Accuracy	67.46	68.53	79.74	78.66
Cor.cause prob. (%)	(61.25, 35.65)	(60.90, 35.95)	(72.73, 29.16)	(71.92, 31.20)
Cor.diagn.prob. (%)	(88.93, 10.19)	(87.59, 9.17)	(92.21, 8.33)	(90.30, 8.15)
Rank order	(1.61, 1.15)	(1.60, 1.14)	(1.27, 0.63)	(1.30, 0.69)

Table 6.14: BN structures: Description of Experiment 3

Experiment 3	
Model parameters	learnt parameters
Training set	Ntrain= 100 to 5000 cases
Learning algorithm	SEMD and m-est
Test set	sim: Ntest= 1000 cases
Number of sets (n)	10

Figure 6.21: BN structures: Diagnosis accuracy vs.size of training set



Dependency with the size of the training set

Fig.6.21 shows the dependency of the average diagnosis accuracy with the size of the training set for the two BN structures under test (Experiment 3 on Table 6.14). From approximately 500 onwards, the SBM outperforms the Noisy-OR BN, although the differences are not meaningful enough. It should be noticed that the training and test cases have been generated based on the same assumptions as those of the SBM, which can explain the better performance of the SBM structure when the training set size is large.

Sensitivity analysis

Experiment 4 (Table 6.15) studied the sensitivity to imprecise parameters of the diagnosis accuracy obtained with both structures. In Fig.6.22(a), it can be observed that there are no meaningful differences in the behavior of the structures in the presence of inaccurate thresholds. However, in Fig.6.22(b), it can be appreciated that the Noisy-OR structure is more sensitive to imprecise definition of prior probabilities than the SBM. With regard to conditional probabilities of symptoms given causes (Fig.6.22(c)), sensitivity is similar for both structures. Finally, Fig.6.22(d) presents the sensitivity analysis when noise is added to all the parameters (thresh-

Table 6.15: BN structures: Description of Experiment 4

Experiment 4	
Model parameters	reference parameters
Training set	Not applicable
Learning algorithm	Not applicable
Test set	sim: Ntest= 1000 cases
Number of sets (n)	1
Ntimes	50

olds and probabilities). The conclusion from Fig.6.22(d) is that there are not clear benefits of the Noisy-OR structure over the SBM with regard to sensitivity to imprecise parameters.

Summary

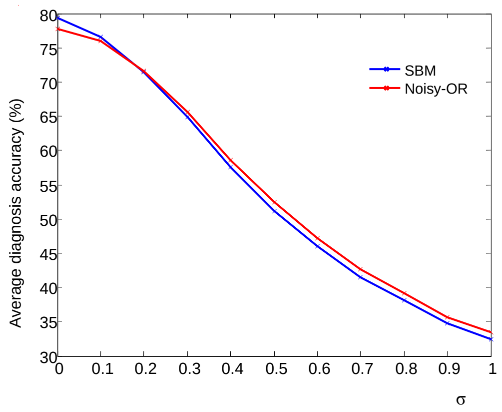
It has been shown that when only a single cause is present (which is normally the case in the domain under study), the SBM and the Noisy-OR structures give similar results, taking into account not only the accuracy, but also their sensitivity to imprecise model parameters. The simplicity of the SBM in comparison with the Noisy-OR leads to the conclusion that the SBM should be preferred for diagnosis in cellular networks. However, in the future, experiments with two simultaneous causes should be carried out to support or reject this statement.

6.4.3 Learning of model parameters

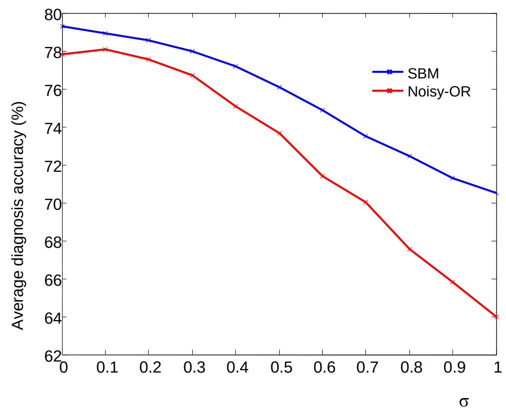
The experiments presented in this section were aimed to evaluate the algorithms presented in Chapter 5 for parameter learning. Those methods are used to calculate the quantitative part of the model, i.e. thresholds for the discretization and probabilities for the BN. The BN was a SBM with two-value symptoms. Therefore, continuous performance indicators were discretized into two states.

The first type of experiments was aimed at testing the working mechanisms of the discretization algorithms. The reduction in complexity of SEMD in comparison with EMD was also quantified. The second class of analysis was focused on comparing the algorithms for probability definition. Finally, the performance of the diagnosis systems obtained after applying the proposed algorithms was studied. The figure of merit was the diagnosis accuracy. In a domain such as cellular networks, where the number of classified cases is very limited, the fact that an algorithm achieves a high performance even with a reduced training set has been considered a quality measure. This is the reason why the behavior of the different methods was analyzed depending on the number of training cases.

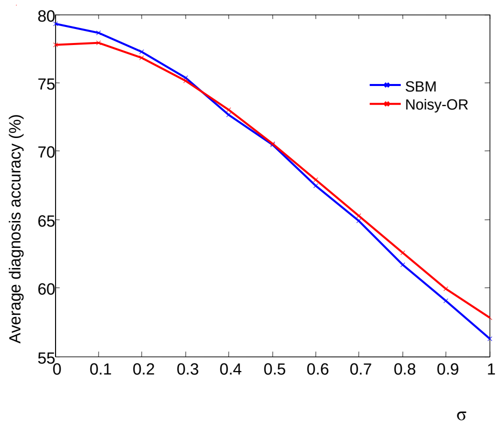
Table 6.16 shows the relations between causes and symptoms assumed for the experiments (e.g. for the SEMD and BMAP algorithms).



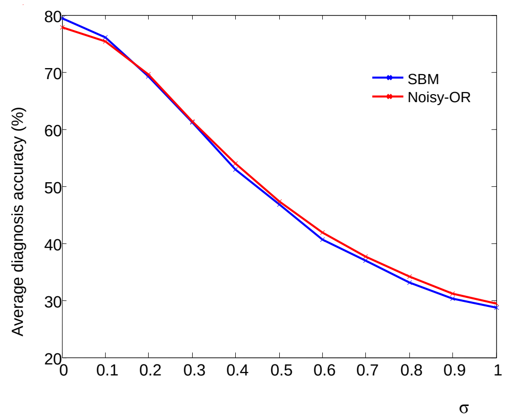
(a) Thresholds



(b) Prior probabilities of causes



(c) Conditional probabilities of symptoms



(d) All parameters

Figure 6.22: Sensitivity analysis to imprecise model parameters for different structures

Table 6.16: Cause - Symptom relations

Cause	**Related symptoms**
UL interference	2, 3, 11, 14, 18, 24, 28, 30, 33, 34
DL interference	2, 3, 12, 13, 19, 29, 31, 33
Lack of coverage	2, 3, 11, 12, 15, 16, 21, 26, 27, 28, 29
HW fault	2, 11, 12, 13, 14, 15, 16, 17, 18, 19, 28, 29
Transmission fault	4, 10
Trasncoder fault	8, 10

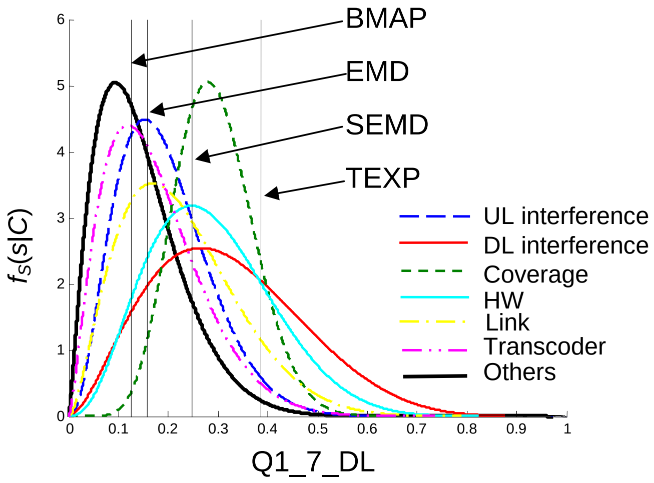


Figure 6.23: Thresholds obtained with different discretization methods

Table 6.17: Average of thresholds (mean, std)

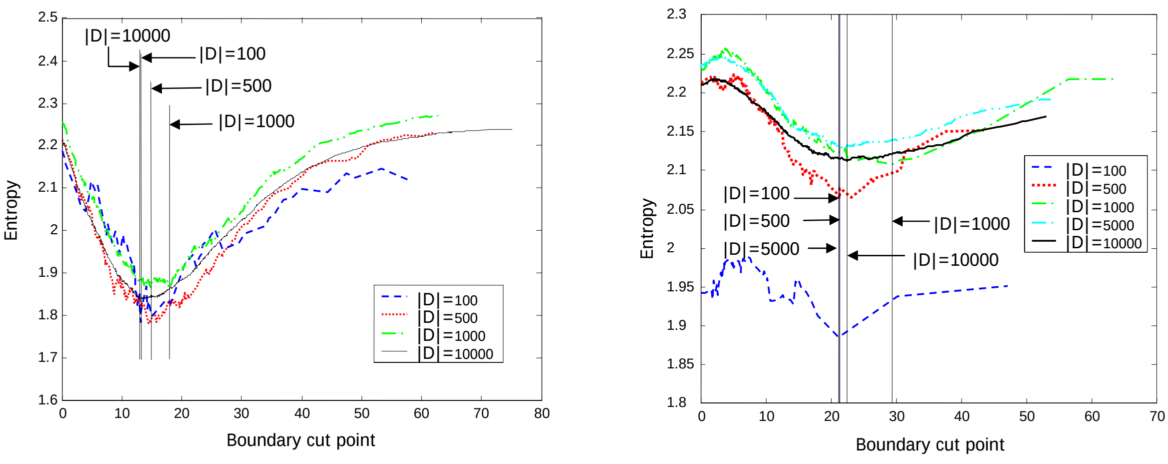
	$\lvert D\rvert = 100$	$\lvert D\rvert = 500$	$\lvert D\rvert = 1000$	$\lvert D\rvert = 5000$
TEXP	17.67	17.67	17.67	17.67
EMD	(9.93, 11.82)	(10.26, 11.15)	(10.08, 10.62)	(9.87, 10.38)
SEMD	(12.57, 13.21)	(11.72, 11.98)	(11.03, 10.92)	(10.39, 10.10)
BMAP	(14.25, 23.25)	(13.96, 16.36)	(13.34, 14.30)	(13.33, 14.17)

Discretization methods

Firstly, the thresholds obtained with the different algorithms proposed in Chapter 5 have been compared. For example, Fig.6.23 shows the theoretical pdfs of the symptom 12, "ratio of downlink quality samples out of band 0", conditioned to the different causes. Thresholds obtained with the different discretization methods are also depicted in that figure. In this example, the number of cases in the training set was 1000. The difference in the threshold values can be clearly appreciated in the figure, although no rule can be followed to select one of the thresholds among the others.

Table 6.17 compares the thresholds obtained with the different discretization methods. The average of thresholds over all symptoms has been calculated using 50 different training sets. In Table 6.17, the average of the average of the thresholds and the average of the standard deviation of the thresholds is shown depending on the number of cases in the training set ($|D|$). It can be observed that the value obtained with SEMD gets closer to that achieved with EMD as the size of the training set is increased.

In order to study the behavior of the entropy methods, the entropy for two symptoms has been depicted versus the boundary cut point (Fig.6.24). It has been calculated as eq.(5.31) according to EMD, for different sizes, $|D|$, of the training set. The minimum value for each curve, which corresponds to the best threshold, is also outlined in the figure. The symptom



(a) Received Uplink Level

(b) A bis component in Drop Call Rate

Figure 6.24: Class entropy versus boundary cut point

in Fig.6.24(a) is the "percentage of samples with uplink signal level lower than -100 dBm" (RXLEV_UL), which is directly related to causes "lack of coverage" and "hardware fault". In Fig.6.24(a), it is worth noting that entropy curves almost coincide, regardless of the different sizes of the training sets. However, for the symptom shown in Fig.6.24(b), "A-bis component in Dropped Call Rate" (A-bis_dcr), curves corresponding to different sizes of the training set are not so close as in Fig.6.24(a). The explanation to those differences is the fact that for the symptom in Fig.6.24(a) the related causes are the faults in approximately half of the cases in the training sets. On the contrary, the only cause directly related to the symptom presented in Fig.6.24(b), "transmission failure", have a prior probability of only 0.04. That means that the number of cases whose cause was "transmission fault" in a small training set is normally very scarce. Entropy curves for SEMD are similar to those in Fig.6.24.

The following experiments compared the complexity of SEMD versus EMD. As described on page 120, the complexity of SEMD is $\frac{2}{K}$ times that of EMD, where K is the number of causes. Furthermore, the complexity of EMD depends on the number of boundary cut points to be evaluated. Table 6.18 presents the number of boundary points (averaged over all the symptoms) for training sets of different sizes. The table also shows the complexity of EMD and SEMD, measured as the number of multiplications in the algorithms. In this example, $K = 7$, thus $\frac{2}{K} = 0.286$. It can be appreciated that the reduction in complexity is large enough to justify the use of SEMD over EMD even in situations where the accuracy achieved with SEMD is slightly lower than that of EMD. This advantage will be more evident for networks with a high number of causes, which is normally the case. A reduced complexity is especially important when thresholds are frequently calculated. If the network is stable and thresholds are computed only occasionally, complexity is not so crucial.

Table 6.18: Complexity of EMD and SEMD

	points	**EMD compl.**	**SEMD compl.**
$\lvert D\rvert = 100$	60	840	240
$\lvert D\rvert = 500$	332	4648	1328
$\lvert D\rvert = 1000$	691	9674	2764
$\lvert D\rvert = 5000$	3436	48104	13744

Probability definition

The aim of the experiments presented hereafter was to compare the probabilities obtained with the different methods proposed in Section 5.5.2. 5000 training cases were the base to obtain a set of thresholds, according to the EMD method, and a set of probabilities, according to MEST. Those probabilities, which were named *standard probabilities*, were compared with the probabilities obtained following each of the algorithms. Probabilities obtained with all methods based on data coincide asymptotically, thus the algorithm selected to calculate the standard probabilities is irrelevant as long as the number of cases is enough. The root mean square error over all the probabilities in the BN was computed as

$$E = \sqrt{\frac{\sum_{k=1}^{M} (t_k - s_k)^2}{M}} \tag{6.5}$$

where M stands for the number of probabilities in the BN, t_k are the probabilities to be compared and s_k are the standard probabilities.

Table 6.19 shows the expected value of E (obtained by doing the average over 50 different training sets) depending on the size of the training set ($|D|$). In this case, $M = 168$, which corresponds to 24 symptoms and 7 conditional probabilities per symptom (due to the 7 causes). In this calculus, for each symptom, only the probability of one of the states was considered, as the probability of the second state is directly derived from the probability of the first one. In table 6.19, it can be observed that for large training sets all methods but the one based on knowledge achieve similar errors in probability estimation. However, for reduced number of training cases, MEST outperforms the other methods. It should be pointed out that the training cases used to calculate the standard probabilities were different to the ones used in the experiments. This explains that the error is different to zero for the MEST method and 5000 training cases.

Figures of merit

The aim of the following experiments (Table 6.20) was to compare the diagnosis accuracy achieved with the different methods proposed in Chapter 5 for probability and threshold calculations. Two scenarios were distinguished depending on the number of available training cases: 50 or 5000 cases.

Table 6.19: Root Mean Square Error for BN probabilities

	$\|D\| = 100$	$\|D\| = 500$	$\|D\| = 1000$	$\|D\| = 5000$
EXP	0.262	0.262	0.262	0.262
MLE	0.175	0.070	0.052	0.029
MEST	0.129	0.065	0.049	0.028
BDF	0.373	0.079	0.046	0.028

Table 6.20: Learning algorithms: Description of experiments 1-4

Experiment 1	
Model parameters	learnt parameters
Training set	Ntrain=50 cases
Learning algorithm	all
Test set	network cases
Number of sets (n)	500
Experiment 2	
Model parameters	learnt parameters
Training set	Ntrain=5000 cases
Learning algorithm	all
Test set	network cases
Number of sets (n)	500
Experiment 3	
Model parameters	learnt parameters
Training set	Ntrain=50 cases
Learning algorithm	all
Test set	sim: Ntest= 1000 cases
Number of sets (n)	500
Experiment 4	
Model parameters	learnt parameters
Training set	Ntrain=5000 cases
Learning algorithm	all
Test set	sim: Ntest= 1000 cases
Number of sets (n)	500

Experiments 1 and 2 used network cases as test set. In Experiment 1, the performance of the systems was studied when the number of training cases was scarce (Fig.6.25). On the contrary, Experiment 2 used large training sets (Fig.6.26). In the figures, it can be observed that the best results are achieved when the MEST algorithm is applied to estimate the BN probabilities, specially when the number of training cases is scarce. For a large number of training cases (Fig.6.26), all methods based on training data presents a similar behaviour. In addition, the standard deviation of the diagnosis accuracy obtained with the MEST method is smaller than those achieved with the other methods.

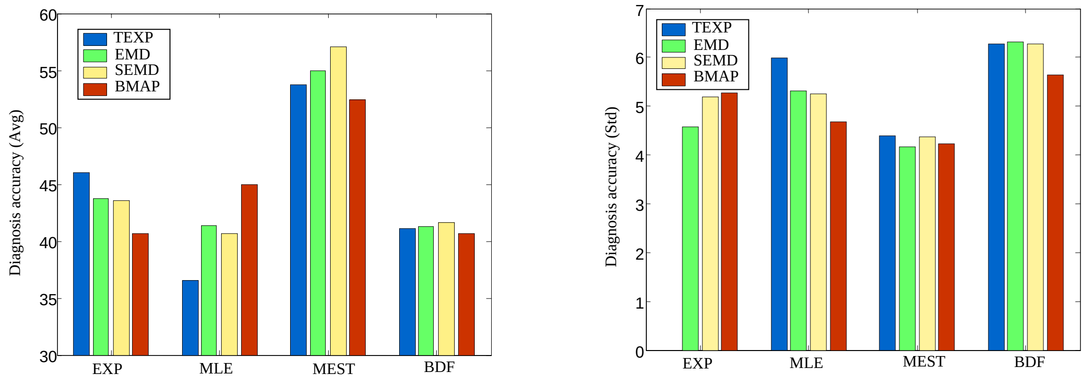
Regarding the discretization methods, differences among the methods are in most cases negligible. When the number of training cases is scarce (Fig.6.25), it can be observed that the ranking of the discretization method depends on the applied algorithm for probability calculation. For a large number of training cases (Fig.6.26) the data-based discretization methods outperform the method based on human expertise. Furthermore, EMD and SEMD are slightly better than BMAP. However, the standard deviation obtained with BMAP is the lowest one.

In Experiments 3 and 4, simulated cases were used as test sets. Results are presented in Fig.6.27 and Fig.6.28 for small and large training sets, respectively. In Fig.6.27, it can be observed that the best performance was obtained when probabilities were estimated applying the MEST method, regardless of the discretization technique, followed by BDF. On the contrary, when probabilities were elicited by experts or obtained applying the MLE method, the error was higher. In addition, in Fig.6.27(a) it can also be appreciated that SEMD outperforms the other discretization algorithms when used in combination with MEST. In order to test the statistical significance of the previous results a binomial test [168] was applied. The hypothesis was that the best results were achieved when the SEMD method was used to set the thresholds and the MEST algorithm was applied to define the probabilities in the BN. When the MEST algorithm was used, SEMD proved to outperform the other discretization methods (at least at the 90% confidence level) in 74% of the experiments for EMD, in 58.6% of the experiments for BMAP and in 80.4% of the experiments for TEXP.

For large training sets (Experiment 4), the achieved diagnosis accuracy was similar for all methods but for those based only on human expertise, which presented higher diagnosis error (Fig.6.28).

Dependency with the size of the training set

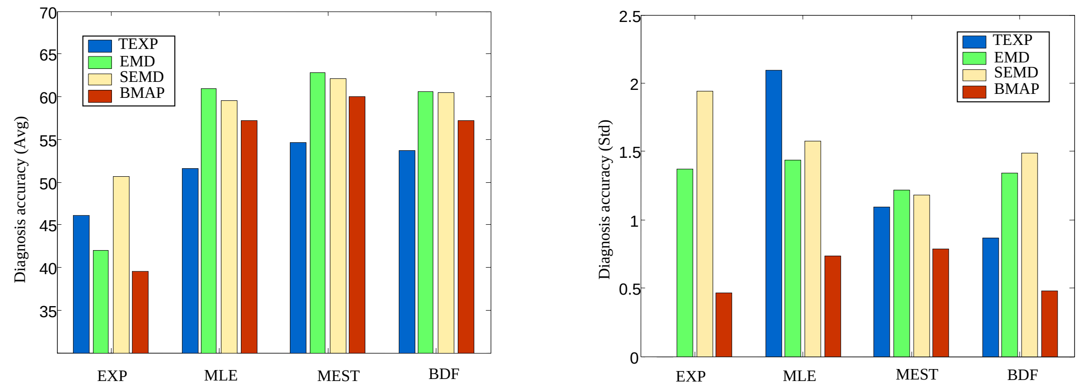
The dependency of the results on the size of the training set was also studied (Table 6.21). Fig.6.29(a) (Experiment 5) shows the diagnosis accuracy for the different discretization techniques when MEST was applied to obtain the probabilities. It can be clearly appreciated in this figure that data-based techniques outperform the knowledge-based method. In addition, for most training set sizes SEMD slightly outperforms EMD and BMAP. Accuracies achieved with BMAP and SEMD are very similar, except for low number of cases. Fig.6.29(b) (Experiment 6) presents the diagnosis accuracy calculated using SEMD thresholds for the diverse methods proposed for probability definition. Probabilities elicited by experts provide considerable worse



(a) Average

(b) Standard Deviation

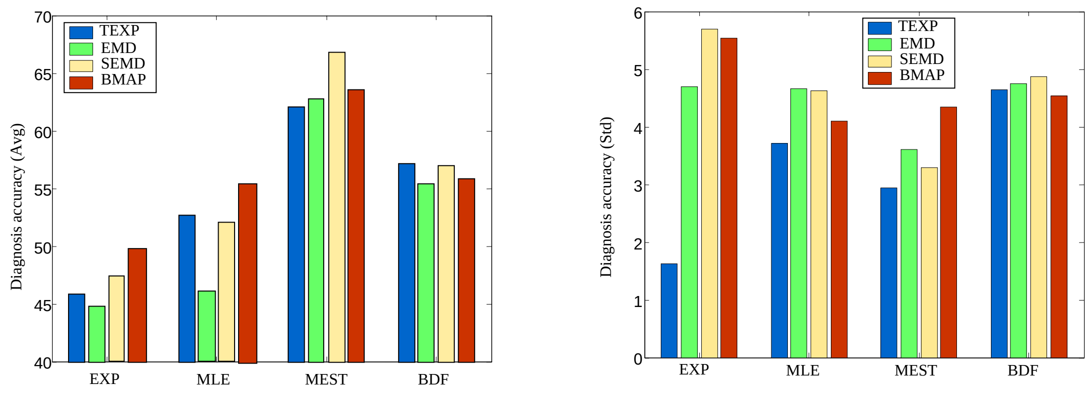
Figure 6.25: Diagnosis accuracy, network cases (training set size=50)



(a) Average

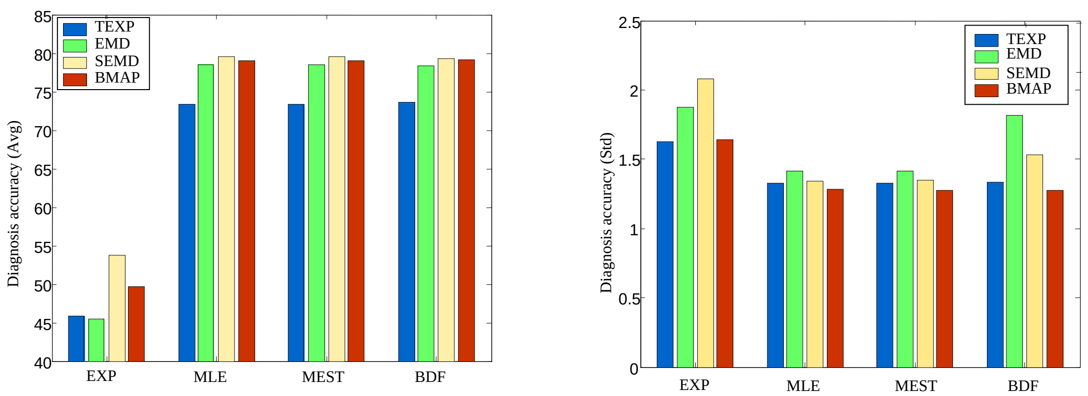
(b) Standard Deviation

Figure 6.26: Diagnosis accuracy, network cases (training set size=5000)



(a) Average (b) Standard Deviation

Figure 6.27: Diagnosis accuracy, simulated cases (training set size=50)

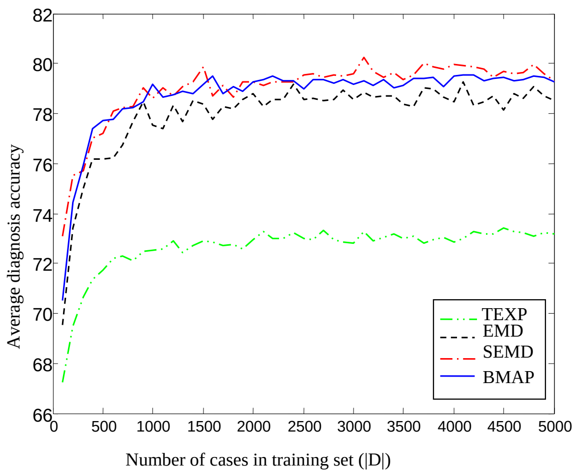


(a) Average (b) Standard Deviation

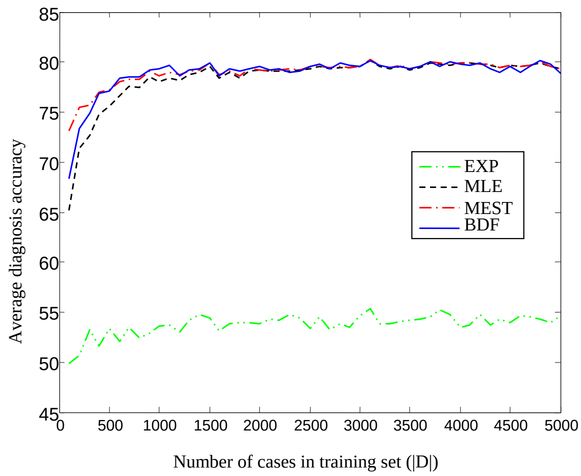
Figure 6.28: Diagnosis accuracy, simulated cases (training set size=5000)

Table 6.21: Learning algorithms: Description of experiments 5 and 6

Experiment 5	
Model parameters	learnt parameters
Training set	Ntrain=100 to 5000 cases
Learning algorithm	MEST and all discretization alg.
Test set	sim: Ntest= 1000 cases
Number of sets (n)	10
Experiment 6	
Model parameters	learnt parameters
Training set	Ntrain=100 to 5000 cases
Learning algorithm	all probability alg. and SEMD
Test set	sim: Ntest= 1000 cases
Number of sets (n)	10



(a) Probabilities computed according to m-estimate



(b) Thresholds computed according to SEMD

Figure 6.29: Diagnosis accuracy vs. size of training set

results than those calculated based on data. For large sizes of the training set (i.e. from 2000 cases onwards) it can be observed that MLE, MEST and BDF give similar results. Nevertheless, for lower number of cases, MEST offers the best accuracy, followed by BDF.

Summary

From the previous experiments, it can be concluded that the best diagnosis performance is obtained by using methods based on data or on a combination of data and expertise. Furthermore, the algorithm used for probability definition has proven to be more important than the method used to calculate the thresholds.

When the number of cases in the training set is large, similar results are obtained with all methods except for those based only on knowledge. In that case, the criterion to select the system should be based on the simplicity of the solution. Therefore, when large training sets are available we propose to use MEST or MLE, in combination with SEMD. On the contrary, when the number of training cases is scarce, the MEST method (followed by the BDF algorithm) has shown the best performance to define probabilities. Among the discretization techniques, SEMD was outstanding when combined with MEST for small training sets. Due to the fact that it is difficult to determine from which size on a training set is considered large, we propose to use the SEMD method for setting thresholds and the MEST algorithm for computing the probabilities, regardless of the size of the training set.

In the extreme situation in which no data is available, the BMAP and BDF methods can still be applied if the parameters a and b of the beta pdfs have been previously estimated by diagnosis experts.

6.4.4 Prevention of imprecision in model parameters

Experimental design

The objective of the study on SBNs and MUI networks was two-fold: to compare their performance, with different degrees of smoothness and different belief mapping functions, and to evaluate their sensitivity to imprecise parameter setting.

Firstly, a discrete BN whose probabilities were elicited by experts (Table A.2) was built. The thresholds were obtained from 5000 training cases using the SEMD method. The reason for choosing data-based thresholds instead of knowledge based thresholds was that the thresholds elicited by experts were very inaccurate because a feedback procedure was never done on the real network. Secondly, several SBNs were designed based on the previous discrete BN. Likewise, several networks following the MUI method were also defined.

An example

As an example to better understand the behavior of SBNs and MUIs, let's consider a case whose real fault cause was interference in the downlink path. This case was diagnosed by a 2-state BN, two SBNs (trapezoidal and rect-gaussian functions) and a 3-state MUI BN. In all cases, the parameter p was set to 10%. The effect of changing the threshold of a single symptom, RXLEV_DL, was studied.

In this case, the value of this symptom was 7.5% and the threshold was 6.4%, therefore in the case of the 2-state BN, the symptom was in the second state. In the case of the MUI BN, the symptom was in the middle state. The 2-state BN provided a probability of 34.31% for HW fault and 34.29% for interference in DL. Although both probabilities are extremely close, the cause was erroneously diagnosed as HW fault, following the criterion of selecting the cause with the highest probability. A lack of coverage was incorrectly diagnosed by the MUI BN, whereas both SBNs achieved a correct diagnostic. Fig.6.30 depicts the probability of the three causes (lack of coverage (Cov), interference in the DL path (DL int.) and hardware fault (HW)) vs. the value of the threshold for the RXLEV_DL symptom. The different shapes of the transition band between states for each system can be clearly appreciated in the figure.

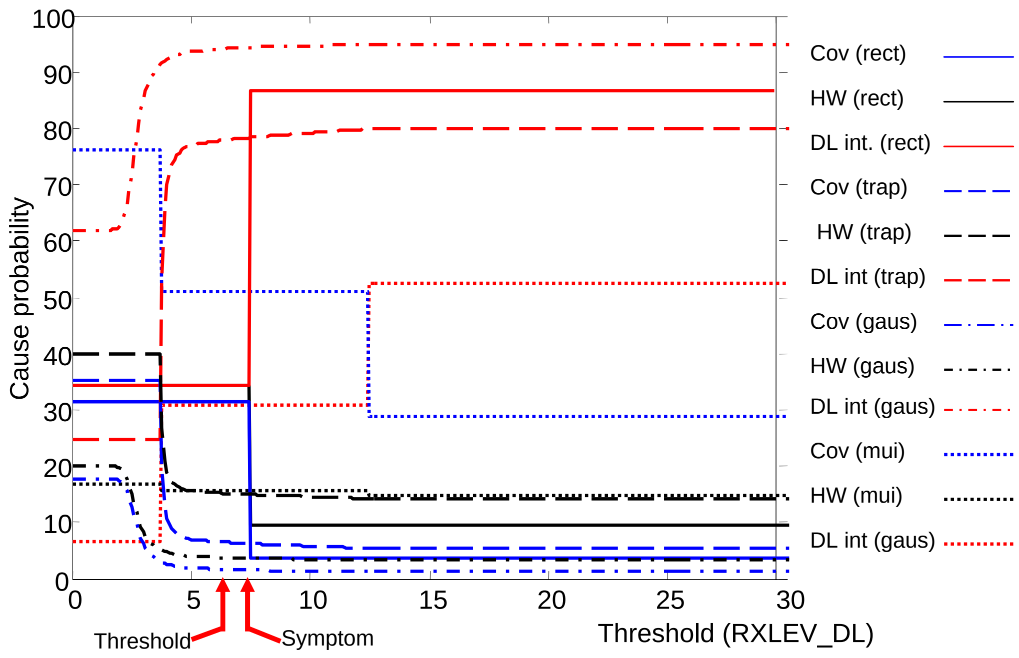


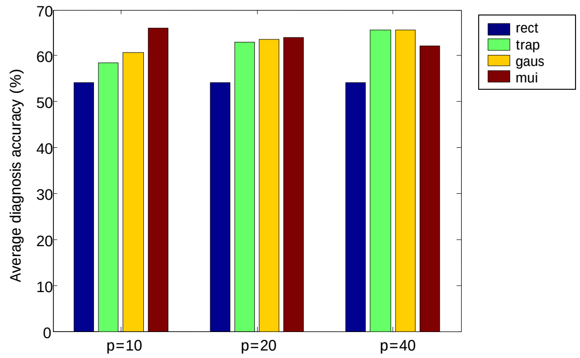
Figure 6.30: Cause probability for a single case vs. threshold value

Figures of merit

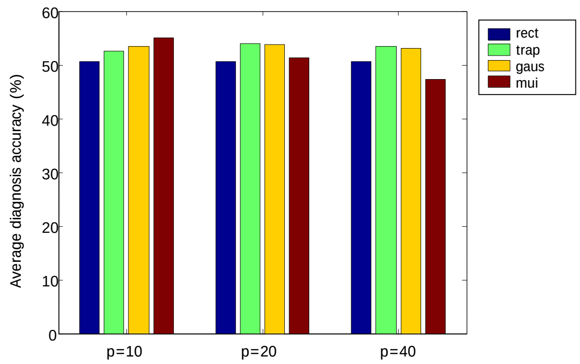
The aim of these experiments (Table 6.22) was to compare the performance of a discrete BN to various SBNs and MUI networks. Table 6.23 shows the average and standard deviation of the achieved diagnosis accuracy for the 2-state BN, the SBNs and the MUI networks in Experiment 1 (simulated cases). Fig.6.31(a) depicts the average diagnosis accuracy. It can be observed that in all cases, the proposed methods outperforms the traditional BN (*rect*). Another remarkable aspect is that for a degree of smoothness of $p = 10\%$, the rect-gaussian function (*gaus*) provides a higher accuracy than the trapezoidal function (*trap*), which can be explained by the smoother shape of the *gaus* function in comparison to the *trap* function. However, as p increases, the gain in accuracy of *gaus* in comparison to *trap* decreases. For $p = 10\%$, MUI outperforms SBNs. As p increases, the accuracy achieved with MUI decreases. Finally, for $p = 40\%$, SBNs outperform MUI. It can be concluded that the diagnostic performance of BNs can be improved by the proposed techniques, both SBNs and MUIs, but the gain depends on the mapping functions and the degree of smoothness. The results of the analysis (Experiment 2 in Table 6.22) on cases from the live network (Fig. 6.31(b)) are similar.

Impact of the degree of smoothness

In the previous experiments, it was observed that the performance depended on the degree of smoothness, p. Experiment 3 (Table 6.24) aims to measure the influence of p on the diagnosis results. Fig.6.32 shows the average diagnosis accuracy for rectangular (2-state BN), trapezoidal and rect-gaussian belief mapping functions vs. p. For traditional BNs, the parameter p does not



(a) Simulated cases



(b) Network cases

Figure 6.31: Average diagnosis accuracy for SBNs

Table 6.22: SBNs and MUIs: Description of experiments 1 and 2

Experiment 1	
Model thresholds	learnt thresholds
Training set	Ntrain=5000 cases
Learning algorithm	SEMD
Model probabilities	expert
Test set	sim: Ntest= 1000 cases
Number of sets (n)	50
Experiment 2	
Model thresholds	learnt thresholds
Training set	Ntrain=5000 cases
Learning algorithm	SEMD
Model probabilities	expert
Test set	network cases
Number of sets (n)	50

Table 6.23: Diagnosis accuracy of methods to overcome parameter imprecision, simulated cases

	p=0 (rect)	**p=10**	**p=20**	**p=40**
trap	(54.11, 1.63)	(58.53, 1.61)	(62.88, 1.55)	(65.52, 1.45)
gaus	(54.11, 1.63)	(60.79, 1.55)	(63.51, 1.47)	(65.66, 1.41)
MUI	(54.11, 1.63)	(65.94, 1.41)	(63.93, 1.49)	(62.17, 1.50)

Table 6.24: SBNs and MUIs: Description of Experiment 3

Experiment 3	
Model thresholds	learnt thresholds
Training set	Ntrain=5000 cases
Learning algorithm	SEMD
Model probabilities	expert
Test set	sim: Ntest= 1000 cases
Number of sets (n)	10

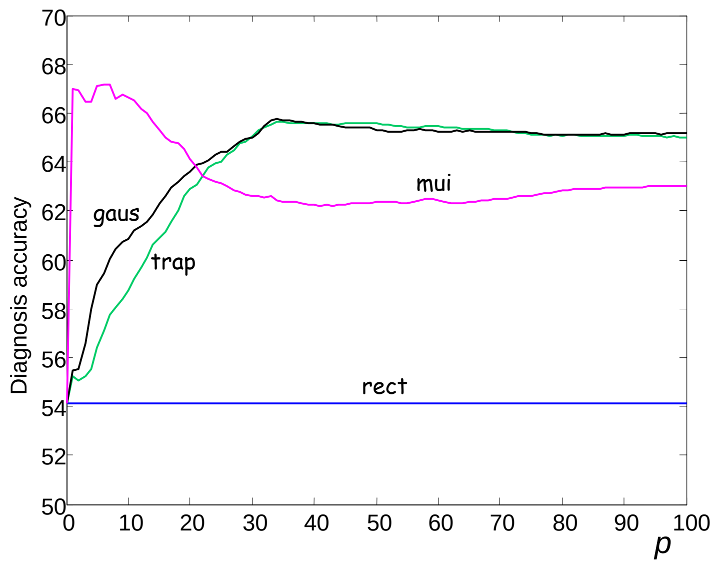
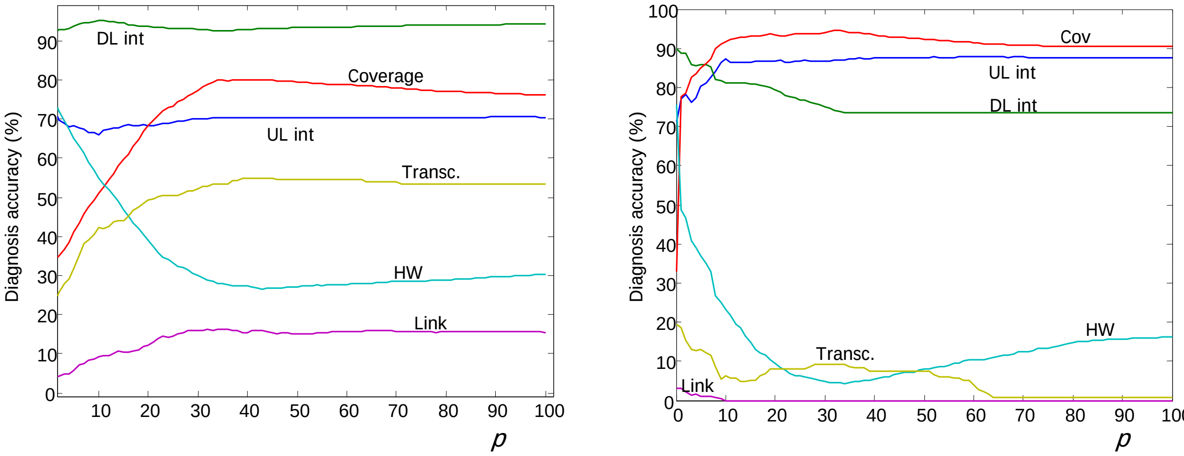


Figure 6.32: Diagnosis accuracy vs. degree of smoothness

exist, which explains the constant value of the accuracy for *rect* functions. For SBNs and MUI BNs, it can be observed that the performance is improved in comparison to the traditional BN, regardless of the p value. For SBNs, the accuracy does not increase further from approximately $p = 35\%$ onwards. This can be explained by the limitation of the p parameter. For MUI, p should not be too high because the accuracy decreases from approximately $p = 10\%$. A gain of up to 13% can be achieved by using SBNs or MUI BNs instead of the 2-state BN.

Fig.6.33(a) shows the diagnosis accuracy per cause vs. the p value for trapezoidal functions. For rect-gaussian functions, the behavior is similar. The influence of p is different for each cause, the total behavior being dominated by the most common cause (i.e. lack of coverage). For example, for the cause "lack of coverage" an improvement of 47% in the accuracy is obtained when p is changed from $p = 0\%$ to $p = 40\%$. Nevertheless, in that case, the accuracy decreases 48% for the cause "faulty hardware". Fig.6.33(b) shows the diagnosis accuracy per cause vs. the p value for a MUI BN.

The type-I error per cause for the 2-state BN is presented in Fig.6.34. From Fig.6.33(a) and 6.34, it can be noticed that those causes that were in most cases misclassified by the 2-state BN, i.e. lack of coverage, link and transcoder failure, are those whose accuracy increases the most



(a) Diagnosis accuracy per cause vs. p for trap function

(b) Diagnosis accuracy per cause vs. p for MUI network

Figure 6.33: Diagnosis accuracy per cause vs. p for MUI

by using SBNs. In Fig.6.33(b), that relation between the gain in accuracy per cause and the type-I error of the 2-state BN is not so clear for MUIs.

Sensitivity analysis

A model with reference parameters (Table 6.25) was taken as the starting point for the analysis of sensitivity to imprecision in model parameters. Those parameters were selected because they were considered to be the "optimum", whereas the added noise represented the imprecision in those parameters. In this analysis it is assumed that the probabilities are defined independently of the thresholds. Thus, the followed methodology consists in fixing the parameters of one type (either thresholds or probabilities) and adding noise to the other parameters. This method is useful for knowledge-based models, in which parameters are independently specified[6].

The sensitivity was separately analyzed for imprecision in the thresholds, in the prior probabilities of causes and in the conditional probabilities of symptoms. In addition, analysis for imprecision in all the parameters at the same time was also performed.

In Experiment 4, sensitivity analysis was carried out for a 2-state BN. Fig.6.35 shows the average diagnosis accuracy vs. the level of noise σ. It can be clearly appreciated that the system is almost unsensitive to imprecision in the prior probabilities. On the contrary, the highest sensitivity is related to imprecision in thresholds.

In Experiment 5, the behavior of BNs, SBNs and MUI networks in the presence of imprecise parameters was compared (Fig.6.36). It can be observed that with no noise ($\sigma = 0$), the best results are obtained with the 2-state BN. This is logical as the parameters in the model were calculated to be the optimum for that model. SBNs and MUI networks introduce uncertainty

[6]On the contrary, in the algorithm described in Chapter 5 for parameter learning based on training data, thresholds are firstly defined. Afterwards, probabilities are calculated as functions of the thresholds.

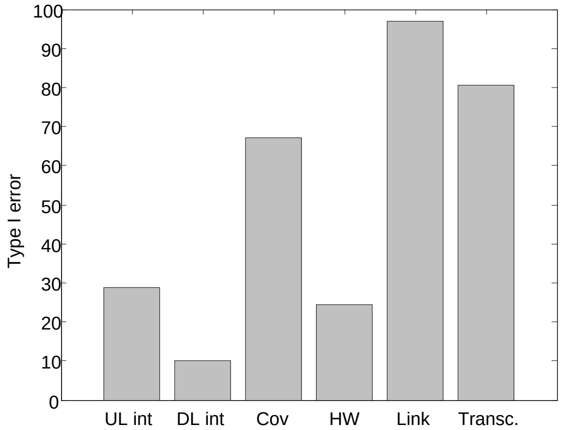


Figure 6.34: Type I error for 2-state BN

Table 6.25: SBNs and MUIs: Description of experiments 4 and 5

Experiments 4 & 5	
Model parameters	reference parameters
Training set	Not applicable
Learning algorithm	Not applicable
Test set	sim: Ntest= 1000 cases
Number of sets (n)	1
Ntimes	50

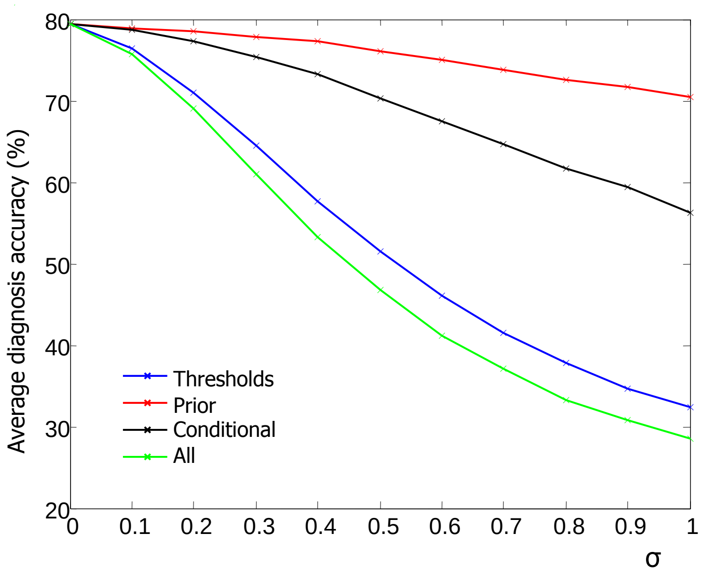


Figure 6.35: Imprecision in model parameters for a 2-state BN

about the state of the symptoms. In other words, they equalize the probabilities of the different states. The consequence is that the performance is worst with SBNs and MUI networks than with the 2-state BN. However, as imprecision is added to the parameters, SBNs and MUI networks outperform the 2-state BN. In Fig.6.36(a), it can be clearly observed that the 2-state BN (*rect*) is the most sensitive to imprecision in the thresholds. Even with low level of noise, SBNs and MUI networks outperform the 2-state BN. In addition, in the figure it can also be appreciated that SBNs are less sensitive to imprecision in threshold definition than MUI networks. SBNs with rect-gaussian functions outperform SBNs with trapezoidal functions in presence of imprecision. Regarding prior probabilities of causes (Fig.6.36(b)), all networks behave in a similar way. When only conditional probabilities are changed (Fig.6.36(c)), MUI networks achieve the best diagnosis accuracy for medium and high levels of noise, whereas there are no meaningful differences among the others. Fig.6.36(d) presents the realistic situation in which all parameters present some degree of imprecision. It can be noticed that the behavior is dominated by the imprecision in thresholds. Therefore, in the presence of uncertainty the best results are obtained with SBNs with rect-gaussian functions, followed by SBNs with trapezoidal functions. The worst diagnosis accuracy corresponds to traditional BNs. Experiments were repeated for the network cases. The obtained results are the same ones as those presented above for the simulated cases.

Finally, the influence of the degree of smoothness on the sensitivity to imprecise parameters was analyzed. Fig.6.37 presents the sensitivity to changes in all the parameters for a SBN with different degrees of smoothness.

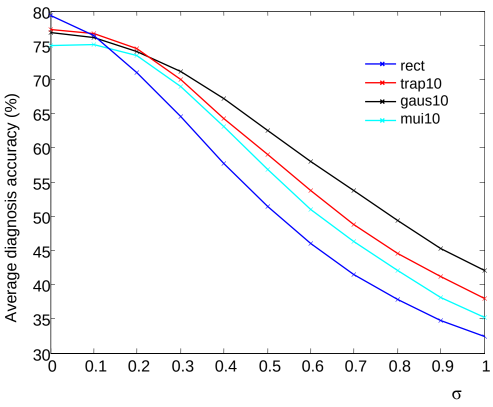
Summary

SBNs has proven to outperform traditional BNs when there is certain degree of inaccuracy in the model parameters. However, when the parameters are calculated based on large databases of training data, BNs are preferred. Among the analyzed functions, rect-gaussian are less sensitive than trapezoidal functions to imprecise definition of parameters. The performance of SBNs is influenced by the selected degree of smoothness. Thus, it is recommended to test the system in the network before regular operation in order to choose the best possible p.

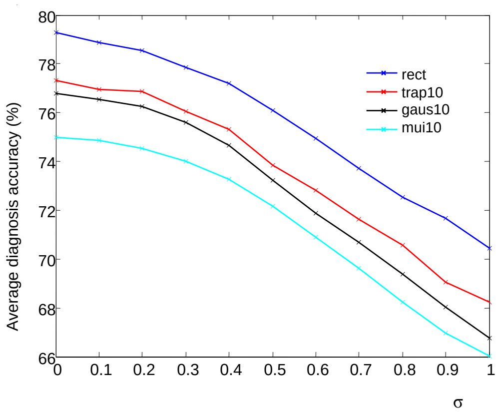
6.5 Discussion

In this chapter, it has been shown that each diagnosis systems proposed in this thesis has some advantages and disadvantages. The main conclusions are summarized in Fig.6.38-6.40. In those figures, the learning algorithms for BNs have been SEMD/mest.

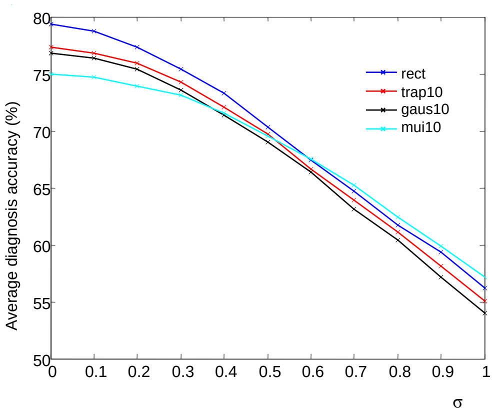
The BC gives very good results when the parameters are good, in the sense that they are fitted to the cellular network under study. Suitable parameters can be obtained if a large database of training cases is available. In Fig.6.38, it can be observed that for a training set of 5000 cases, the accuracy obtained with the BC is much higher than that obtained with the other diagnosis systems. However, when the model parameters are inaccurate, the performance



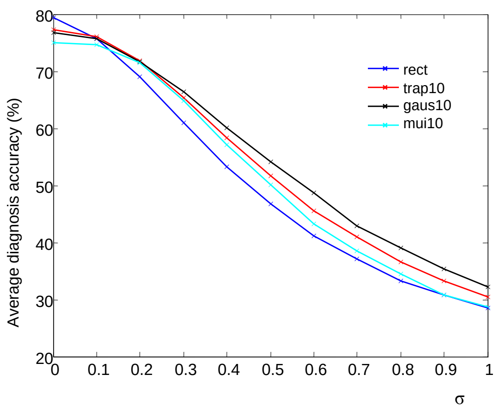
(a) Thresholds



(b) Prior probabilities of causes



(c) Conditional probabilities of symptoms



(d) All parameters

Figure 6.36: Sensitivity analysis to imprecise model parameters for SBNs and MUI

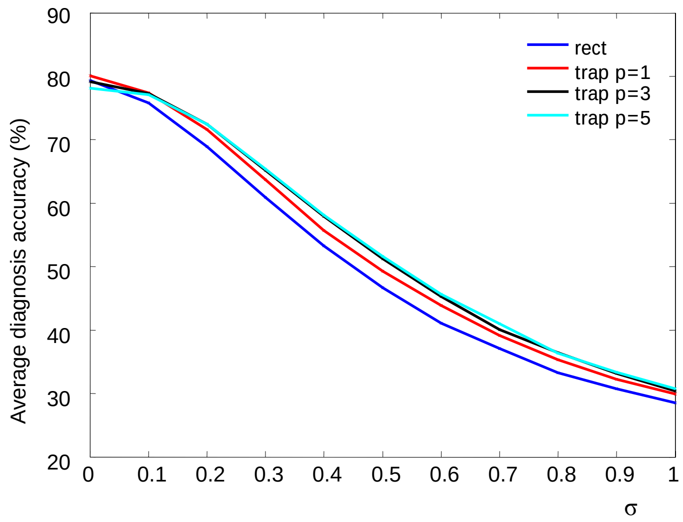


Figure 6.37: Sensitivity analysis to imprecise model parameters for different degree of smoothness

of the BC is highly degraded, as also shown in Fig.6.39 and 6.40. On the one hand, inaccurate parameters can be due to a small training set. In Fig.6.40, it can be seen the steep slope of the diagnosis accuracy of the BC when the size of the training set is reduced. On the other hand, inaccuracy can be also present in knowledge-based systems. As shown in Fig.6.39, the sensitivity of the BC to imprecise parameters is very high in comparison to the other proposed systems.

Reasonable results can be obtained with BNs for medium and small sizes of training sets. The SBM structure is preferred to the Noisy-OR structure because diagnosis accuracy is similar for both structures and the SBM is much easier to implement. In Fig.6.38, it can be observed that for a small training set ($N = 50$), the SBM outperforms the BC. This can be also noticed in Fig.6.40, since from a certain size downwards the accuracy achieved with the SBM is higher than that obtained with the BC. In addition, the behavior of BNs is more independent on the training set size than that of the BC. On the contrary, results obtained with the SBM for large training sets are much worse than those obtained with the BC.

Two different tests have been performed with BNs. The first one (knowledge-based BN), depicted in green in Fig.6.39, is the results of the experiment presented in green in Fig.6.35. In this test, it has been considered that thresholds and probabilities are both defined by experts. Thus, noise is independently added to thresholds and probabilities, simulating the inaccuracy in both types of parameters. On the contrary, the curve in red in Fig.6.39 (data-based BN) analyzes the sensitivity of the accuracy to imprecise parameters for data-based systems. In that case, according to the methods proposed in Chapter 5, thresholds are first calculated and then probabilities are computed depending on the value of the thresholds. Thus, in this test, firstly the thresholds are determined by means of any discretization method. Secondly, noise is added to the thresholds. Subsequently, probabilities are computed depending on the noisy thresholds. Finally, noise is added to the probabilities. In this case, as expected, the accuracy

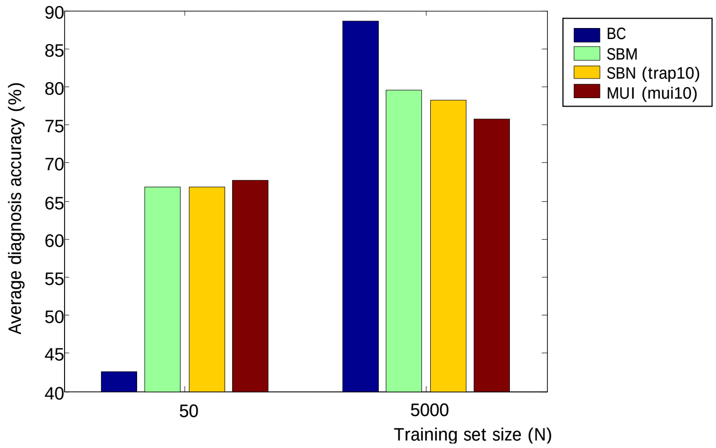


Figure 6.38: Comparison diagnosis systems

is less sensitive to inaccurate parameters than in knowledge-based BNs. This is due to the fact that probabilities are calculated for the actual thresholds. In Fig.6.39, it can be observed that in any case the SBM is less sensitive to imprecise parameters than the BC.

Finally, SBNs and MUI networks improve the accuracy obtained with knowledge-based BNs when there is certain degree of imprecision in the parameters. For example, in Fig.6.39, the smaller sensitivity of SBNs to imprecise parameters compared to knowledge-based BNs or BCs can be observed.

In summary, some general recommendations extracted from these conclusions are the following:

- When there is a large number of training examples available, the BC should be used as diagnosis system.
- When the number of training examples is scarce, the diagnosis system should be based on a BN with a SBM structure. The parameters of the model should be calculated using the SEMD and mest algorithms.
- When there are no training cases, a BN should be built using the knowledge of diagnosis experts. If the model parameter are not precise, which is normally the case, SBNs or MUI networks should be used.

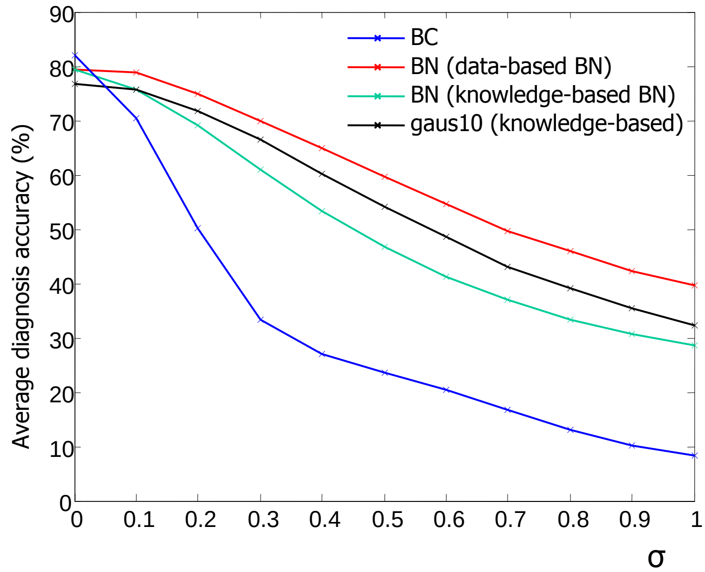


Figure 6.39: Sensitivity of diagnosis systems to imprecise parameters

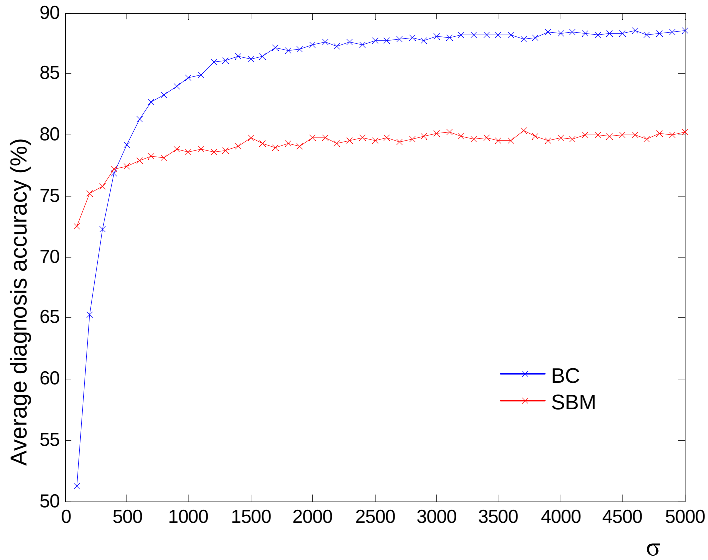


Figure 6.40: Dependency of diagnosis systems on the size of the training set

Chapter 7

Conclusions

This chapter summarises the research carried out and the methods developed, describes the main difficulties encountered and assesses the most important conclusions obtained in this thesis. In addition, the main contributions are summarized and some future lines of work are suggested.

7.1 Results

In this thesis, a framework for automated fault diagnosis in cellular networks has been established. The systems proposed in this thesis are intended to improve operational efficiency in current and future networks, by means of the following benefits:

- Automated troubleshooting (TS) reduces the time required to identify the fault cause in cells with problems. This means that network performance is enhanced as the downtime and the time with reduced quality of service is significantly limited. Consequently, the number of unsatisfied users decreases and the threat posed by other cellular network operators and other emerging technologies is counteracted.
- Thanks to the automation of TS procedures, operational costs are reduced because fewer personnel with a lower skill level is required. This is due to the fact that the majority of problems can be rectified with the help of the automated tool. The knowledge of highly experienced staff, which is released from this work, can be utilized for other aspects of network optimization, thereby further increasing network performance.
- One of the main difficulties encountered in this thesis has been the lack of documentation about fault management in cellular networks. The steps followed to diagnose the fault cause in a cell are normally known only by a few experts who acquired that knowledge from hand-on experience. Hence, one additional benefit of the methods proposed in this thesis is that the TS knowledge can be stored in the automatic tool, therefore making the company less vulnerable to staff fluctuation.

Apart from the main "scientific" conclusions that were detailed in Section 6.5, it is important to stress those engineering aspects learnt from working with cellular network operators. This may help anyone who attempt to create a similar system for cellular networks or any other application domain in which a database of cases is not available. The most important issues to take into account when building an automatic troubleshooting system are the following:

- A close cooperation between network operators, manufacturers of network equipment and academic researchers has proved to be essential for the success of the project. On the one hand, the contribution from network operators was the expertise in the application domain (i.e. troubleshooting in the RAN of cellular networks). On the other hand, the researcher contributed with her knowledge about artificial intelligence and cellular networks. The manufacturer of network equipment was responsible of implementing prototype tools and of the negotiations with the operators. Finally, the important task of bringing together the theory on automatic systems and the expertise in TS was performed by the network operator together with the researcher.

- When creating a knowledge-based model, a feedback cycle is required in order to fine-tune the model parameters:

  1. Firstly, the artificial intelligence technique should be selected by the knowledge engineers[1]. If BNs are chosen, then the structure of the BN should be defined (e.g. SBM).
  2. Then, the model parameters should be provided by the experts in the application domain (e.g. experts in fault management in the RAN) with the help of the knowledge engineers or an automatic tool. This stage, despite being conceptually simple, is one of the most difficult ones in model building. The reason is that TS experts are used to a language completely different to the terms utilised by the knowledge engineers. Consequently, it is recommended that knowledge engineers have also a background in the domain under study (mobile communications in this case), so that knowledge transfer can be simplified. In addition, the nature of the tasks performed by TS experts prevented them from spending sufficient time building a model. This is specially true if they do not see the immediate benefits of having an automatic system. In this phase, it is important to reduce the number of parameters that experts have to specify.
  3. Once the model is defined, it has to be tested in a cellular network. That implies that a prototype TS tool should be available at this stage. The model should be evaluated on some faulty cells and its results should be compared with the diagnosis supplied by an expert. If the diagnosis of the automatic tool is wrong, then the TS expert together with the knowledge engineers should analyse which parameters of the model should be changed so that the expected results are obtained. This procedure should be repeated as many times as required. This can be a very time-consuming process.

[1]The knowledge engineer is a person, but could also be a computer program, that interprets the information presented by experts in the application domain and defines a model based on that information.

Thus, it requires a strong commitment from the operator so that experts do not give up after analysing a few cases.

In conclusion, apart from the theoretical contributions, the importance of this thesis lies in bringing together two completely separated worlds, such as theoretical research and operational network management. For this task, it has been essential to cooperate with several parties, each one with their own interests and ways to solve problems.

The relevance of the research is emphasized by the interest and involvement shown by some cellular network operators, like Telefónica or France Telecom, and manufacturers such as Nokia. In addition, as a result of the work carried out in the agreement with Nokia and in the Gandalf project [3, 2], a commercial troubleshooting tool, Moltsen TheCure [5], has been deployed.

7.2 General appraisal

From the very beginning, the research had a focused and easy to describe target: to design, create and test an automated fault diagnosis system for the radio access networks of cellular systems. Despite the simple target it proved far more difficult to accomplish such task. Along the development of the thesis the main problems that have had to be overcome have been the following:

- The first difficulty came from the multidisciplinary nature of the thesis: these fields were too wide to go in great depth into every topic related to the thesis. On the contrary, most of the research in telecommunication engineering is focused on a much narrower field, which requires looking into the ultimate mathematical or physical explanation of things. Because of that, specially at the beginning of the thesis, the feeling of not knowing enough of everything was very frustrating. Very often the main difficulty was knowing when to stop doing research into a specific topic. Regarding mobile communication systems, studying causes and symptoms that may happen in current cellular networks can be as complex as desired. For example, the study of faults can either stop by knowing that a type of fault is a hardware fault or the research can continue looking into which are the pieces of equipment that can suffer a hardware fault. Another example: a symptom could be bad quality in the uplink. Alternatively, it would have been possible to go one step further and define bad quality as the number of samples out of quality band 0. Or the formula for the symptom could be even defined from counters in the Network Management System. Regarding artificial intelligence, the same challenge than those with mobile communications arose. For example, in the area of BNs itself, many researchers are focusing on specific topics, e.g. inference algorithms. Another example is discretization of continuous variables, which is a main issue in machine learning. Therefore, an effort was required to look at the thesis from a *system* point of view, without going into too many details in every single topic.

- Another problem in the development of this thesis was the lack of related bibliography. Although there are many books and journals related to current cellular networks (e.g.GSM, GPRS, etc.), they are more focused on theoretical research in communications rather than research regarding the operation of the network. This is due to several reasons. Firstly, it seems that the main interests of the academic community are long-term research and simulation-based studies regarding the theoretical performance of cellular networks. Thus, network management is considered as a secondary topic since it is more related to current networks rather than networks that will be deployed in the future. However, for equipment manufacturers and network operators, network operation is one of the main concerns as it is a main cost contributor. Although, they invest many resources in network management, all activities related to the operation of their networks are normally considered highly confidential. In addition, due to the urgent nature of operation tasks, quite often operation procedures are not documented and they are only known by those members of the staff working in this area. Consequently, very few references could be found on symptoms (Key Performance Indicators and alarms), on base station equipment or on how diagnosis is done in cellular networks. Thus, although the beginning of the thesis was a survey on the state of the art, there were very few publications on the subject and the research had to be conducted by interviewing network operators and manufacturers. In this sense, the collaborations with Nokia in the first place and France Telecom at the end was very important. On the contrary, references on diagnosis in other application fields were numerous. Hence, considerable effort was required to extract the information relevant for the thesis from the varied existing documents on diagnosis and artificial intelligence.

- Another consequence of the confidentiality of network operation was that evaluating the systems proposed in the thesis was extremely difficult. In fact, the thesis was facing huge challenges when the evaluation of the proposed methods had to be initiated. On the one hand, there was not any Commercial off-the-shelf (COTS) software which simulated the problems and symptoms required to test diagnosis systems in cellular networks. In addition, designing a new simulator with the demanded accuracy would require a huge effort that could constitute a thesis by itself. On the other hand, obtaining data from live networks was also a barrier because of the reluctance of network operators to provide information from their networks. Even when an operator supplied a database with the values of symptoms in some cells of a real network, this information still proved to be insufficient because no information was given about neither which were the cells experiencing problems nor the causes of those problems. Obtaining six hundred cases from a live network (see page 158) was a milestone in the development of the thesis. Once those data were available, algorithms could be developed to simulate any desired number of cases.

- Related to the two first items (that is the multidisciplinary nature of the thesis, the lack of interest of the academic community in network operation and the “system-approach” followed in the thesis), it has also been very difficult to find proper journals in which to publish the contributions of this thesis. Very often, submitted papers were returned in short without being reviewed because they were out of the scope of the intended journal. Journals focused on mobile communications suggested to send the paper to journals oriented to artificial intelligence, whereas artificial intelligence journals suggested telecommunication publications.

The aforementioned drawbacks are a consequence of the type of practical study developed in this thesis. On the contrary, this applied research provided many satisfactions, such as:

- The work carried out can be considered as pure "engineering", understanding engineering as "the application of scientific and technical knowledge to solve human problems taking into account constraints, such as available resources, technical limitations, flexibility for future modifications, cost and producibility"[32]. In that sense, this type of applied research can be considered a professional fulfilment for any engineer.

- The conclusions of the thesis are more tangible than final results that may be obtained from a more basic research. This is due to the clarity of the aim of the thesis and the medium/short-term results that may be obtained from applying the proposed methodology.

- The chance of working in international projects, with staff from different nationalities, backgrounds and ways of thinking is invaluable. This, undoubtedly, has enriched the work carried out in this thesis.

7.3 Contributions

The contributions presented in this thesis can be grouped in four subsets. A first group consists in the utilization of previously existing techniques in a new application domain and in the compilation of information known by experts in diagnosis, but not reflected in the existing literature:

- Several artificial intelligence techniques (Bayesian Classifier, Bayesian Networks, discretization methods, algorithms to determine probabilities, etc.) have been applied to the diagnosis of faults in the RAN of cellular networks. Those techniques were previously used for the automatic diagnosis in other application fields and for the discretization or probability assessment in the area of machine learning. However, this is the first time that some solutions have been proposed for automatic diagnosis in the RAN of cellular networks.

The second subset of contributions consists in the definition of a system for automatic fault management in cellular networks:

- The architecture of an automatic diagnosis system for the RAN of cellular networks has been proposed, i.e. the main elements and interfaces that the diagnosis system should include and the scenario where it should be located.

- A knowledge base for diagnosis in current cellular networks (i.e. GERAN) has been defined, that is the main causes of excessive dropped call rate in these networks, their symptoms and their related parameters have been identified.

The third subset of contributions comprises models and methods for automatic diagnosis. Those models and methods are not only valid for current cellular networks (i.e. 2G and 2.5G), but they can also be used for any other cellular network, such as 3G or multi-system networks. Most techniques can even be utilized for diagnosis in applications outside the cellular networks domain. In particular, the main contributions are:

- An approximation for the conditional pdfs of symptoms given the causes as beta pdfs has been proposed. This model is based on the fact that pdfs for most symptoms represent belief concerning a relative frequency.

- A BN structure, named Central Bayes Model, has been defined. This structure allows the representation of conditions as parent nodes of the causes, thereby being a more faithful model of the causal relations in diagnosis than the Simple Bayes Model. This model is feasible only if the number of conditions and their states are small.

- A novel discretization method based on entropy minimization, SEMD, has been proposed. SEMD combines experience (i.e. which causes are related to each symptom) with training data. This method has proven to achieve better diagnosis accuracy than EMD, another entropy-based discretization method. In addition, the complexity of SEMD is lower than that of EMD.

- A novel discretization method based on hypothesis testing, BMAP, has been proposed. BMAP, like SEMD, combines experience with training data. This method can also be used when no training data is available if the diagnosis experts specify the a and b parameters of the beta pdfs.

- A method to determine the probabilities in a BN based on beta distributions, BDF, has been proposed. Like BMAP discretization method, BDF may be applied when no training data are available, but in that case diagnosis experts have to specify the a and b parameters of the beta pdfs.

- Two techniques have been proposed in order to prevent inaccuracies due to imprecision in model parameters. One of the main benefits of both techniques is that knowledge acquisition and inference have the same complexity as in conventional BNs. The first method, Smooth Bayesian Networks, applies likelihood evidence to soften the steep shape of the probabilities around the thresholds. In addition, different belief mapping functions were proposed. According to the second method, Multiple Uniform Intervals, 3-state symptoms are built from the 2-state symptoms specified by experts.

- Specific knowledge acquisition procedures for diagnosis in cellular networks have been proposed. The resulting generated model may be based either on a BC or on a discrete BN, according to different structures.

Finally, evaluation of the proposed techniques has been one of the most difficult tasks in this thesis. Because of the lack of databases with classified cases and simulation tools, algorithms had to be designed to simulate diagnosis examples. In addition, methods had to be defined to test and compare the different diagnosis systems. In summary, the main contributions in these areas are:

- An algorithm has been proposed to generate simulated cases based on modelling the conditional probabilities of symptoms as beta pdfs. Simulated cases may be used as training sets or test sets for evaluating new classifiers.
- A method to empirically evaluate the sensitivity of diagnosis systems to imprecision in the discretization of continuous symptoms has been suggested. That method is based on previously existing algorithms for sensitivity analysis to inaccurate probabilities.
- A BC for diagnosis in current cellular networks has been designed and evaluated.
- Different BN structures have been used to implement diagnosis systems for GERAN and they have been tested and compared. In addition, the proposed methods to prevent imprecision in the BN parameters have also been evaluated.
- Different techniques to estimate the probabilities in BNs and to discretize continuous variables have been evaluated and compared.
- Some trials have been carried out in live GSM/GPRS networks. The TS tool has proved to diagnose the fault cause faster than a human expert and with a similar diagnosis accuracy. The obtained results, regarding the time to carry out the diagnosis, the accuracy and the usability of the tool, confirmed the feasibility of the diagnosis system based on BNs and its ability to reduce operational costs.

7.4 Future work

Lines of research that might continue the work in this thesis can be grouped in three main sets:

- Improvement and fine-tuning of the diagnosis model for current cellular networks:
  - More trials in live networks would be helpful to reinforce the practical conclusions of this thesis.
  - An immediate complementary work would be to enlarge the diagnosis model for current cellular networks by including more causes and symptoms.
- Diagnosis model for other RAN technologies: The logical continuation of the performed studies is the application of the techniques proposed in this thesis to another RAN technologies, such as UMTS, WLAN or multi-system networks. Some work in this area has been already started [33, 116, 184], e.g. a basic diagnosis model for UMTS based on a SBM has been defined.
- Diagnosis techniques:

– In this thesis, it has been concluded that the performance of the SBM and the Noisy-OR structures are similar under single fault assumptions. A line of future work is to compare those BN structures when more than a cause takes place at the same time.

– In machine learning literature, several discretization algorithms can be found. In the future, the feasibility of applying some of those methods to mobile communications could be studied. In addition, those techniques could be compared with the ones presented in this thesis.

– In the initial analysis of the state of the art, fuzzy logic was considered as an alternative to BN for diagnosis in mobile networks. Future research may include the application of fuzzy logic to the problem under study.

Appendix A

Parameters of the diagnosis models

Table A.1: Probabilities for causes (knowledge-based system)

Cause	**Probability (%)**
UL interference	10
DL interference	20
Lack of coverage	40
HW fault	20
Transmission fault	4
Transcoder fault	2
Others	4

Table A.2: Probabilities (%) for the symptoms given the causes (knowledge-based system)

Symptom	**UL int.**	**DL int.**	**Cov.**	**HW**	**Transm.**	**Transc.**	**Others**
2	50	50	70	50	10	10	10
3	50	50	50	10	10	10	10
4	10	10	10	10	70	10	10
8	10	10	10	10	10	70	10
10	10	10	10	10	50	50	10
11	70	10	10	50	10	10	10
12	10	85	10	50	10	10	10
13	10	50	10	50	10	10	10
14	50	10	10	50	10	10	10
15	10	10	50	50	10	10	10
16	10	10	70	50	10	10	10
17	10	10	10	50	10	10	10
18	85	10	10	50	10	10	10
19	10	50	10	50	10	10	10
21	10	10	50	10	10	10	10
24	85	10	10	10	10	10	10
26	10	10	70	10	10	10	10
27	10	10	70	10	10	10	10
28	70	10	50	50	10	10	10
29	10	70	50	50	10	10	10
30	85	10	10	10	10	10	10
31	10	70	10	10	10	10	10
33	50	70	10	10	10	10	10
34	70	10	10	10	10	10	10

Table A.3: Thresholds for symptoms (knowledge-based system)

Symptom	**Thresholds (%)**
2	89
3	31
4	13.5
8	1.3
10	2.5
11	24.5
12	38.5
13	26.5
14	10^{-3}
15	15
16	12
17	8
18	1.5
19	9
21	15
24	0.3
26	38
27	13
28	17
29	32.5
30	1.5
31	16.5
33	16
34	1.8

Table A.4: Parameters of beta pdfs for simulated cases (reference parameters) (I)

Symptom		UL int.	DL int.	Cov.	HW	Transm.	Transc.	Others
2	a	5.4306	3.9471	7.7191	8.0295	1.1558	0.5537	0.8269
	b	1.8123	1.9117	2.2631	2.8307	1.5539	1.9582	0.5103
	d	0.008	0	0.0106	0.1048	0	0	0
3	a	1.3858	1.445	1.222	1.5889	1.0848	1.6779	0.4190
	b	7.3786	3.8898	6.7225	7.519	13.1955	15.6346	1.07
	d	0.0325	0.0164	0.03723	0.1429	0.0143	0.1143	0
4	a	1.5459	2.3749	2.7733	2.7654	1.0648	1.4218	1.0718
	b	15.5762	35.2933	32.9503	28.5116	1.2644	7.7415	7.1976
	d	0.0732	0.1393	0.0691	0.2286	0.0143	0	0.3378
8	a	0.4847	2.4559	2.1089	0.5920	0.9930	1.4481	0.4741
	b	15.9425	183.2544	139.149	16.6126	100.442	3.2929	9.4818
	d	0.7480	0.8197	0.76596	0.8571	0.4714	0	0.8771
10	a	1.3033	0	0	0.6031	0.6137	1.0555	0.9368
	b	23.236	0	0	2.6163	3.0861	2.7305	3.9781
	d	0.9431	0.98361	0.9734	0.9429	0.81429	0.2571	0.9582
11	a	1.6651	3.2156	5.7845	3.1113	1.445	2.1004	1.2
	b	5.0552	28.8799	29.7339	14.8263	6.7865	21.7801	15.1027
	d	0	0	0	0	0	0	0
12	a	3.4018	2.771	9.6918	3.6213	2.7047	2.4622	2.324
	b	14.2219	5.877	23.3563	8.9726	9.3413	11.7932	15.4568
	d	0	0	0	0	0	0	0
15	a	1.0318	2.002	4.2407	3.3344	2.2684	1.2533	1.1384
	b	16.0611	19.6085	10.1133	6.6907	12.1349	10.386	14.9349
	d	0.0650	0.0164	0.0106	0.0381	0.1571	0.1143	0
16	a	1.2114	1.5893	6.9328	3.0309	2.4996	1.0227	0.5839
	b	20.7238	37.1456	38.4937	20.7317	30.2699	21.5439	22.1778
	d	0.0650	0	0.0426	0.1333	0.2286	0.1143	0
17	a	1.689	2.1583	2.7689	3.4576	1.3372	1.304	1.2787
	b	34.2384	33.8312	15.9063	12.604	12.7208	19.3065	25.6285
	d	0.0650	0.0164	0	0	0.04286	0.1143	0
18	a	0.6669	1.5893	1.5749	0.8884	0.6363	1.1332	0.5740
	b	8.8056	37.1456	238.5335	94.1812	23.0758	172.2817	78.9243
	d	0.0650	0.0164	0	0.0095	0.0429	0.1143	0
19	a	1.3553	1.297	2.3783	0.9912	1.2127	1.007	0.6194
	b	63.1786	17.586	77.2833	21.1369	40.6486	58.4586	45.0149
	d	0.0650	0.0164	0	0	0.0429	0.1143	0
21	a	0.2416	0.4499	0.3783	0.2416	0	0	0.2584
	b	3.4262	23.9523	3.2474	3.4262	0	0	84.8375
	d	0.8556	0.93023	0.8417	0.8556	1	1	0.6646

Table A.5: Parameters of beta pdfs for simulated cases (reference parameters) (II)

Symptom		**UL int.**	**DL int.**	**Cov.**	**HW**	**Transm.**	**Transc.**	**Others**
24	a	0.2215	0.3723	0.2156	0.2188	0.5187	0.1837	0.2805
	b	2.643	72.1512	11.0709	24.9523	300.6037	13.0358	60.0557
	d	0.4228	0.8197	0.9362	0.9333	0.6571	0.6857	0.89066
26	a	0.4985	0.9444	0.2345	0.4086	0.7447	0.4883	0.61333
	b	306.8087	1578.8287	9.1463	62.1499	1098.852	244.8163	421.2809
	d	0.8537	0.72951	0.4734	0.4762	0.6714	0.6	0.7420
27	a	0	0	0.2853	0	0	0	0
	b	0	0	42.6056	0	0	0	0
	d	0.9756	0.98361	0.9628	0.9810	1	1	0.9975
28	a	1.2802	1.9186	2.7404	2.6133	1.2965	1.3314	1.1458
	b	6.4754	29.0271	22.7595	18.3039	9.4729	11.5048	11.6509
	d	0	0	0	0.0095	0	0	0
29	a	1.008	1.8426	2.1885	1.9296	1.4494	1.5583	1.109
	b	5.221	4.161	4.1006	4.0181	5.0739	4.6463	5.0305
	d	0.0163	0	0	0	0	0	0
30	a	0.4429	1.0376	0.7758	0.7008	0.5668	1.2112	0.6304
	b	4.2687	99.4427	124.459	112.5546	30.4901	118.7717	36.0866
	d	0.0081	0.0328	0.1915	0.2191	0.0143	0.0286	0
31	a	0.6769	1.0295	0.6106	0.6921	0.7704	1.5475	0.7761
	b	14.2244	3.7669	19.0323	12.2089	13.8725	25.7277	8.642
	d	0	0.0082	0.0638	0.0762	0.0143	0.0286	0
33	a	0.6879	1.0388	0.8038	0.8967	0.8178	1.0788	0.4616
	b	6.9336	5.2752	31.1158	22.9801	13.3768	23.6281	10.9265
	d	0.0081	0.0082	0.0372	0.0762	0.0429	0.0571	0
34	a	0.3554	0.6925	0.4971	0.5746	0.5611	1.1766	0.2330
	b	3.2031	87.4934	76.2603	124.2032	24.4758	336.1149	41.6153
	d	0	0.0082	0.0213	0.1048	0	0	0

Table A.6: Parameters of beta pdfs depending on low value samples (reference parameters)

Symptom / Cause		**With low value samples**	**Without low value samples**
RXLEV_UL (15) / Cov.	a	3.6104	4.2407
	b	8.783	10.1133
	d	0	0.0106
RXLEV_UL (15) / HW	a	2.3045	3.3344
	b	4.9691	6.6907
	d	0	0.0381
RXLEV_UL (15) / Transm.	a	1.4057	2.2684
	b	8.6512	12.1349
	d	0.0429	0.1571
RXLEV_DL (16) / Cov.	a	3.1608	6.9328
	b	18.6011	38.4937
	d	0	0.0426
RXLEV_DL (16) / HW	a	1.2807	3.0309
	b	10.3547	20.7317
	d	0	0.1333
RXLEV_DL (16) / Transm.	a	1.2124	2.4996
	b	17.9198	30.2699
	d	0.0429	0.2286

Appendix B

Approximation of the pdfs of the symptoms given the causes

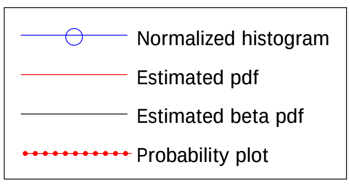


Figure B.1: Legend for figures below (see pag.168)

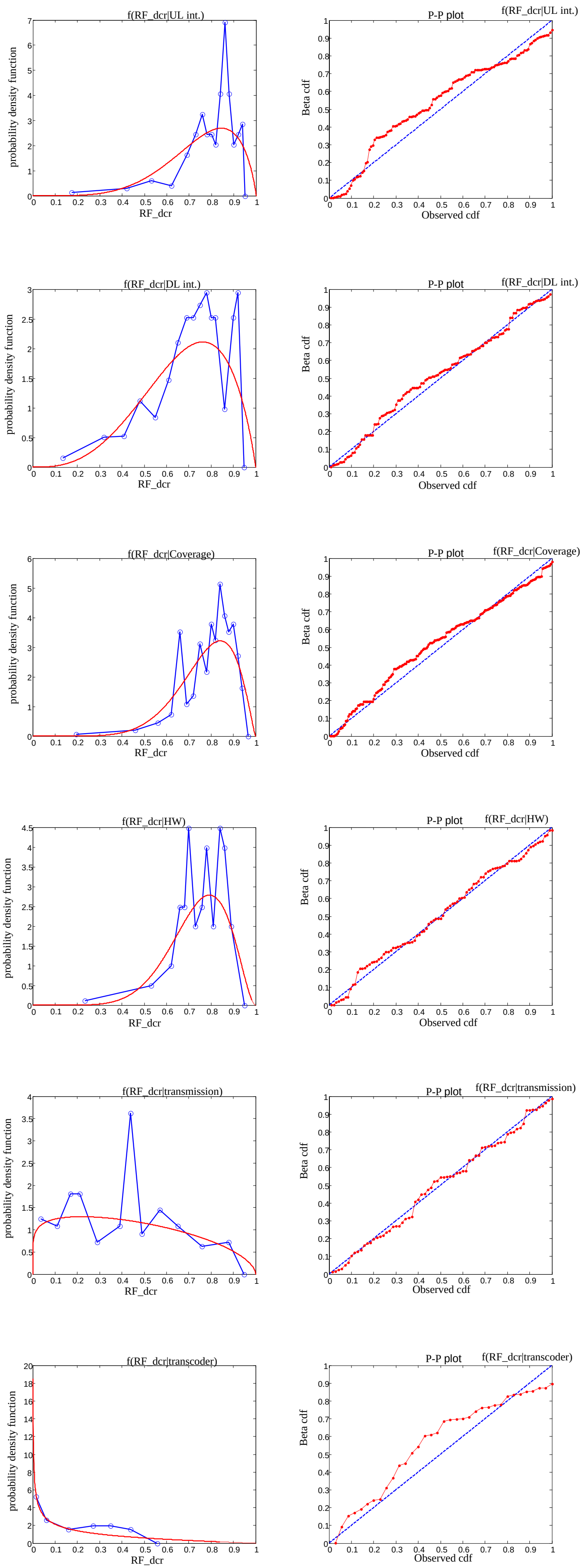


Figure B.2: Modelling of pdfs for symptom RF_dcr

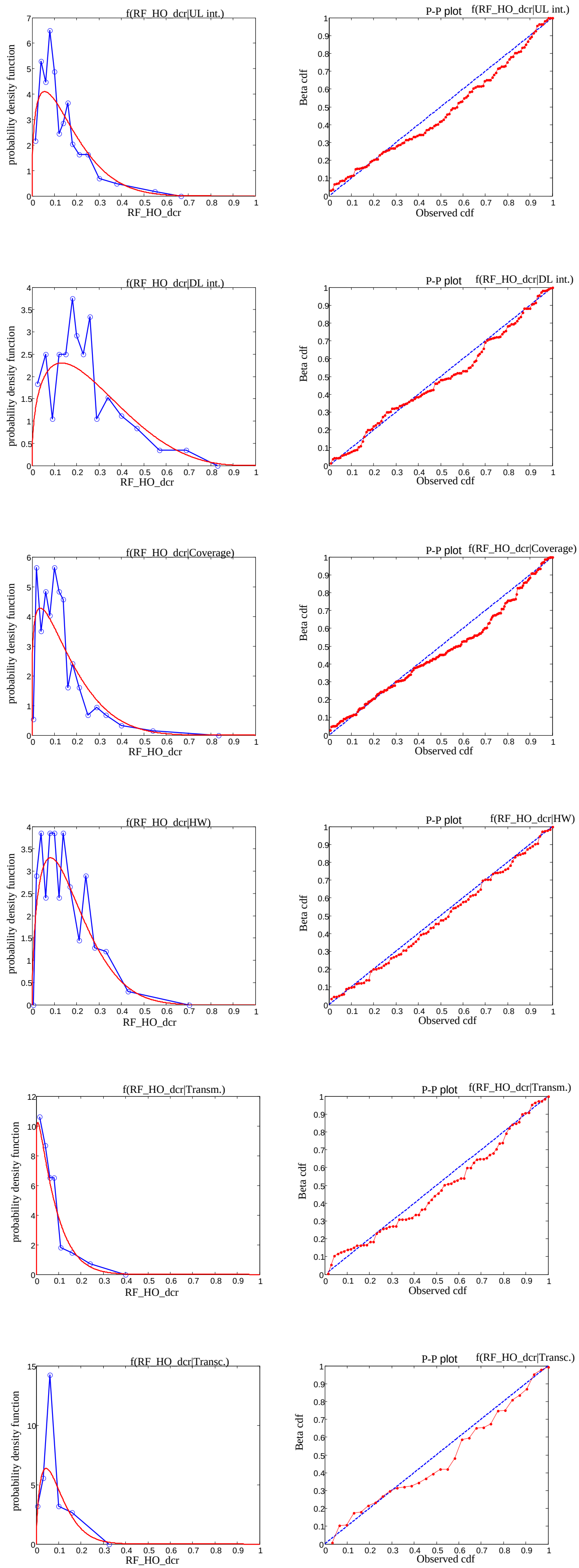


Figure B.3: Modelling of pdfs for symptom RF_HO_dcr

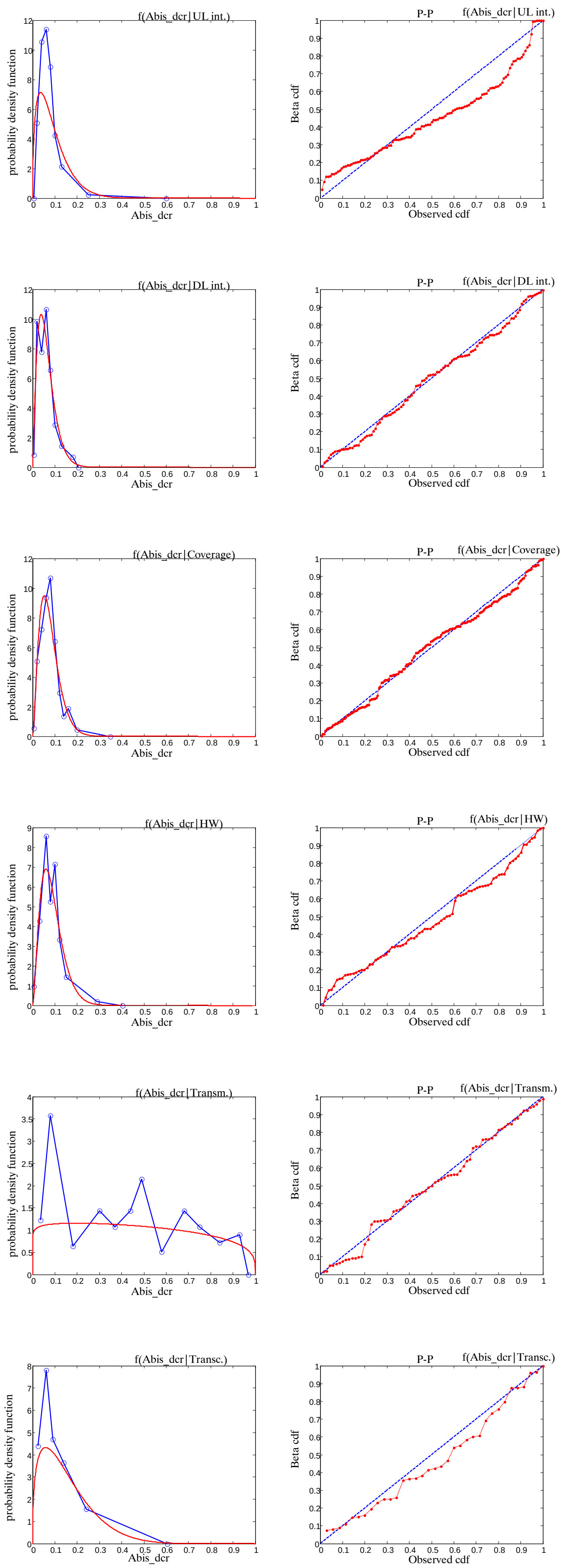


Figure B.4: Modelling of pdfs for symptom Abis_dcr

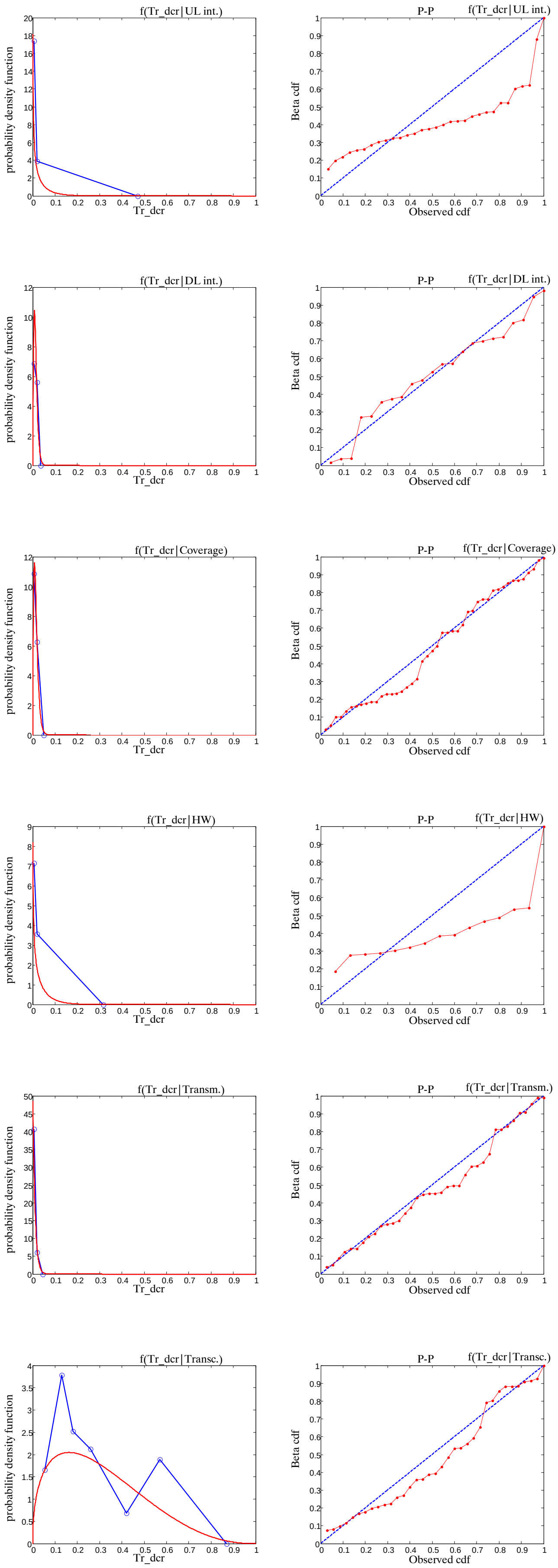


Figure B.5: Modelling of pdfs for symptom Tr_dcr

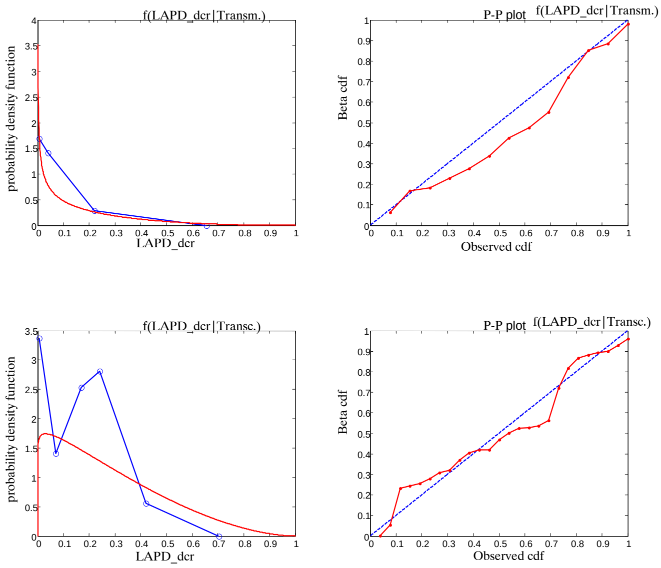


Figure B.6: Modelling of pdfs for symptom LAPD_dcr

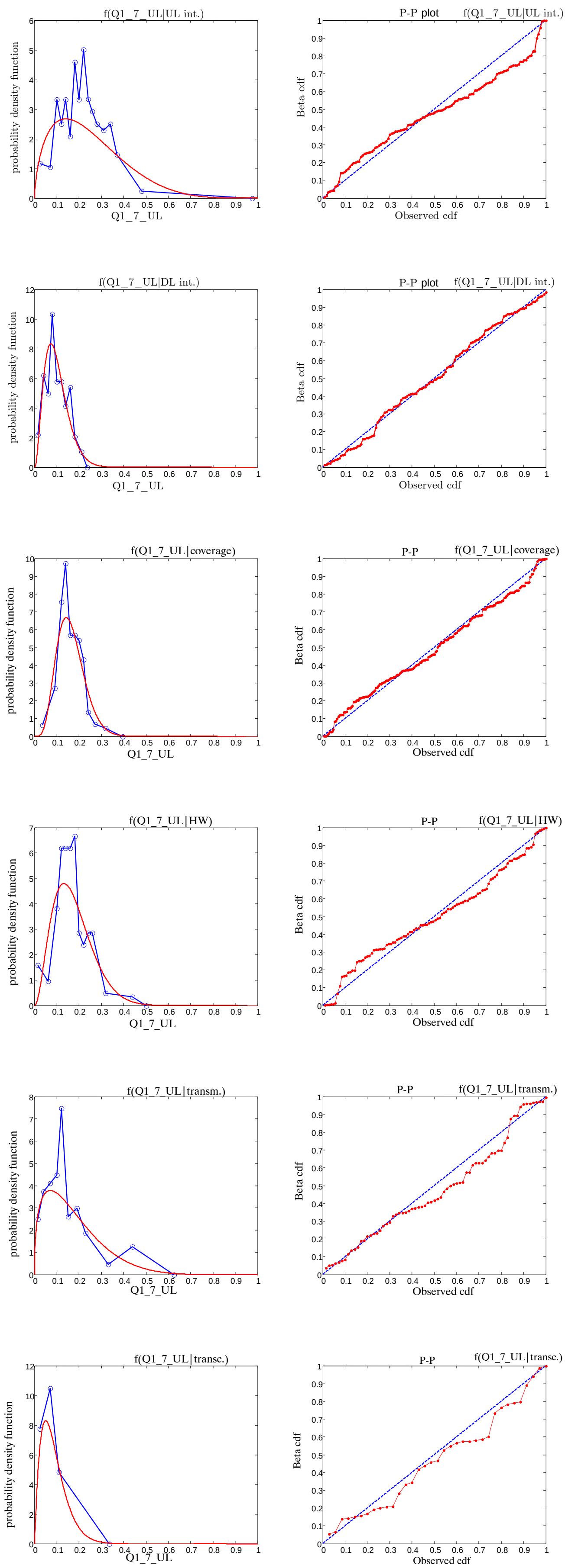


Figure B.7: Modelling of pdfs for symptom Q1_7_UL

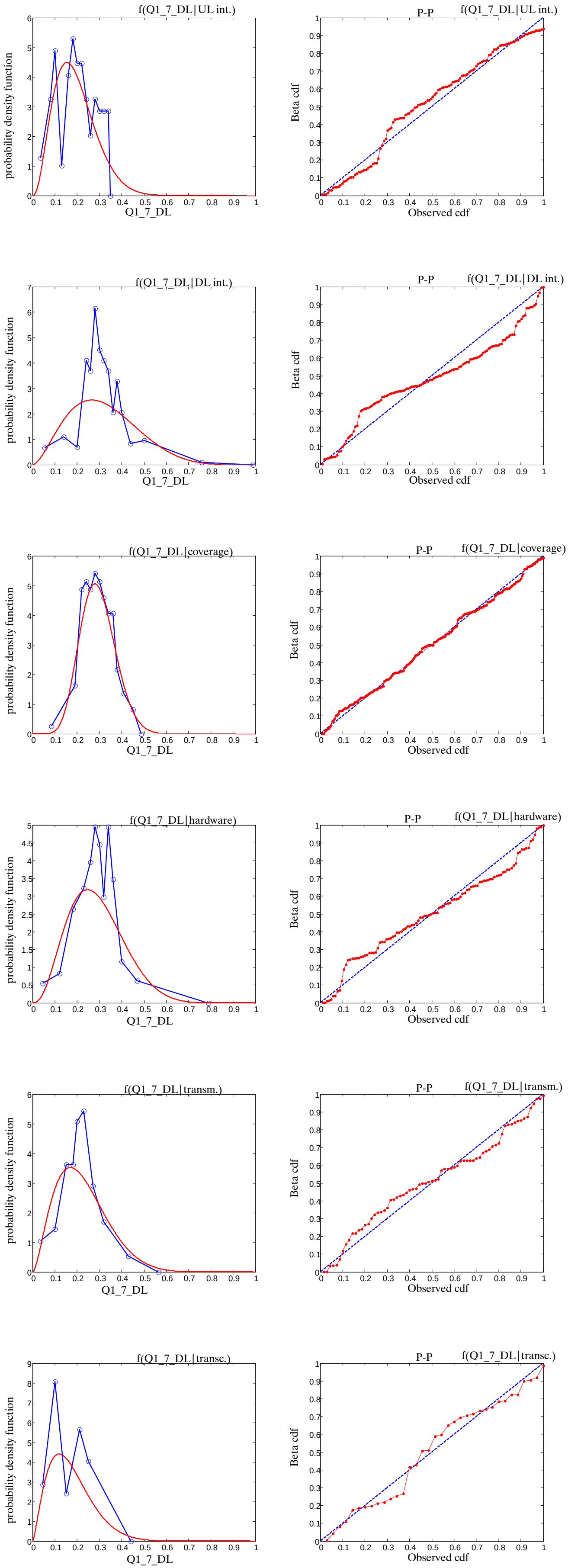


Figure B.8: Modelling of pdfs for symptom Q1_7_DL

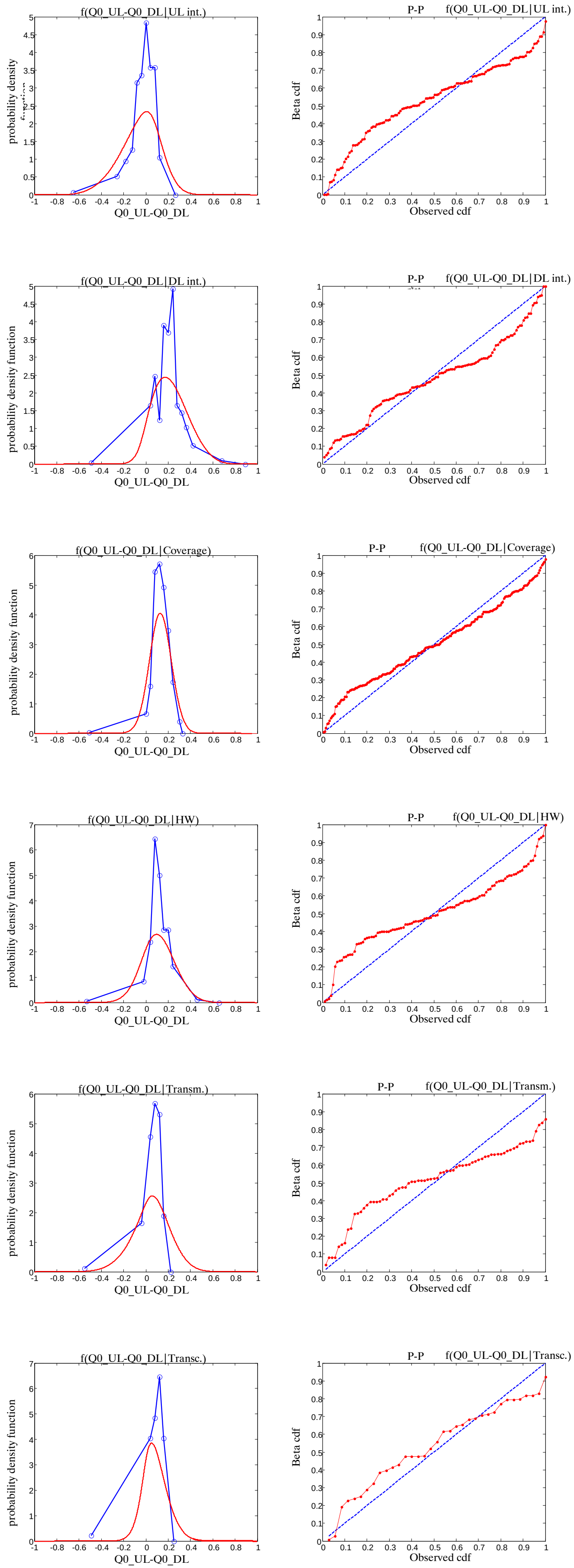


Figure B.9: Modelling of pdfs for symptom Q0_UL-Q0_DL

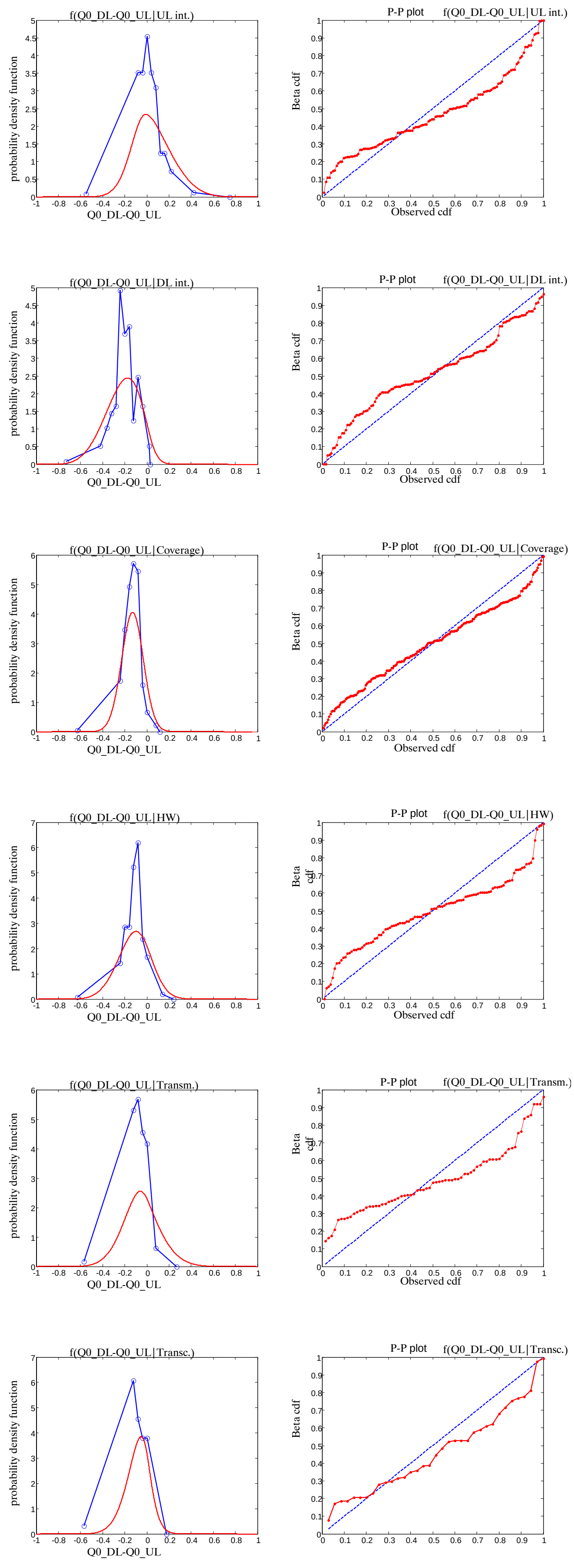


Figure B.10: Modelling of pdfs for symptom Q0_DL-Q0_UL

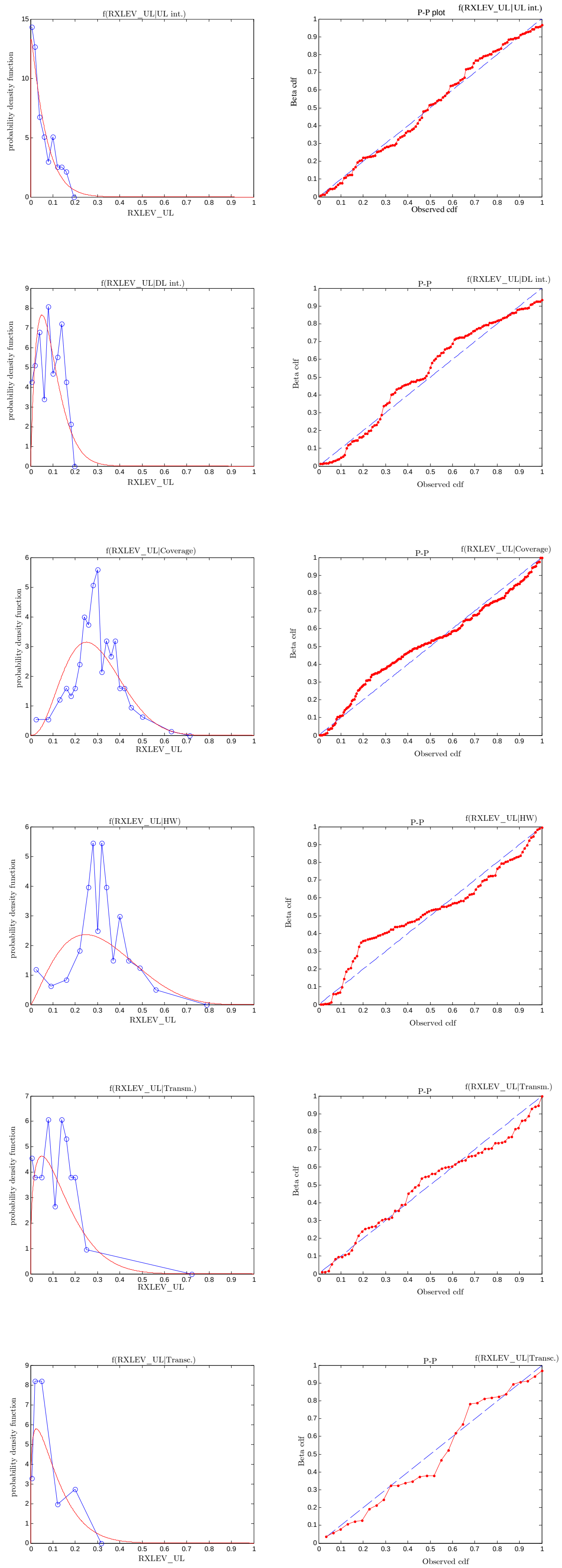


Figure B.11: Modelling of pdfs for symptom RXLEV_UL

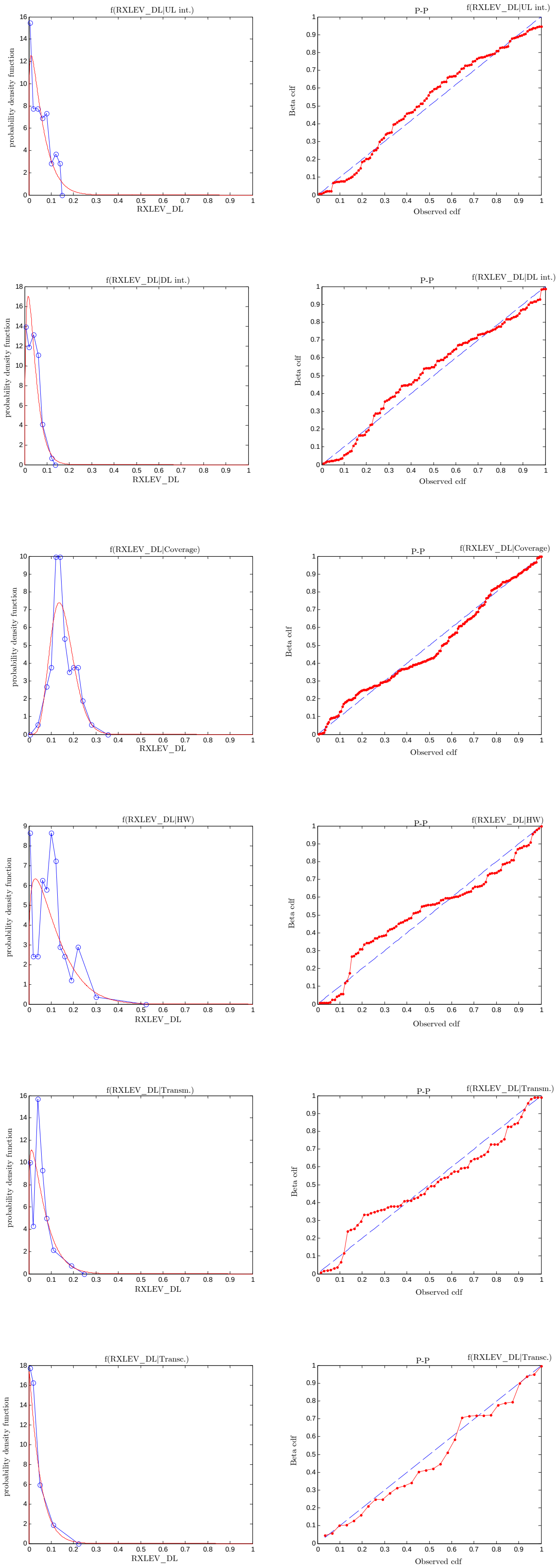


Figure B.12: Modelling of pdfs for symptom RXLEV_DL

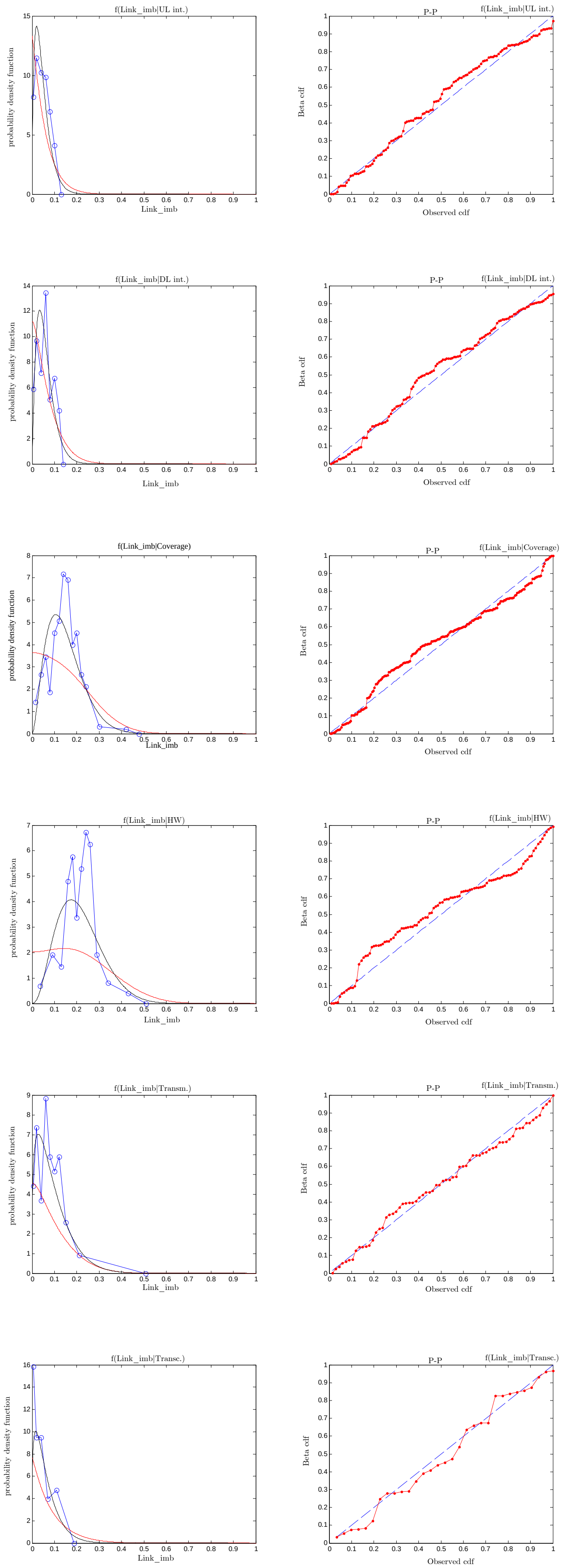


Figure B.13: Modelling of pdfs for symptom "Link_imb"

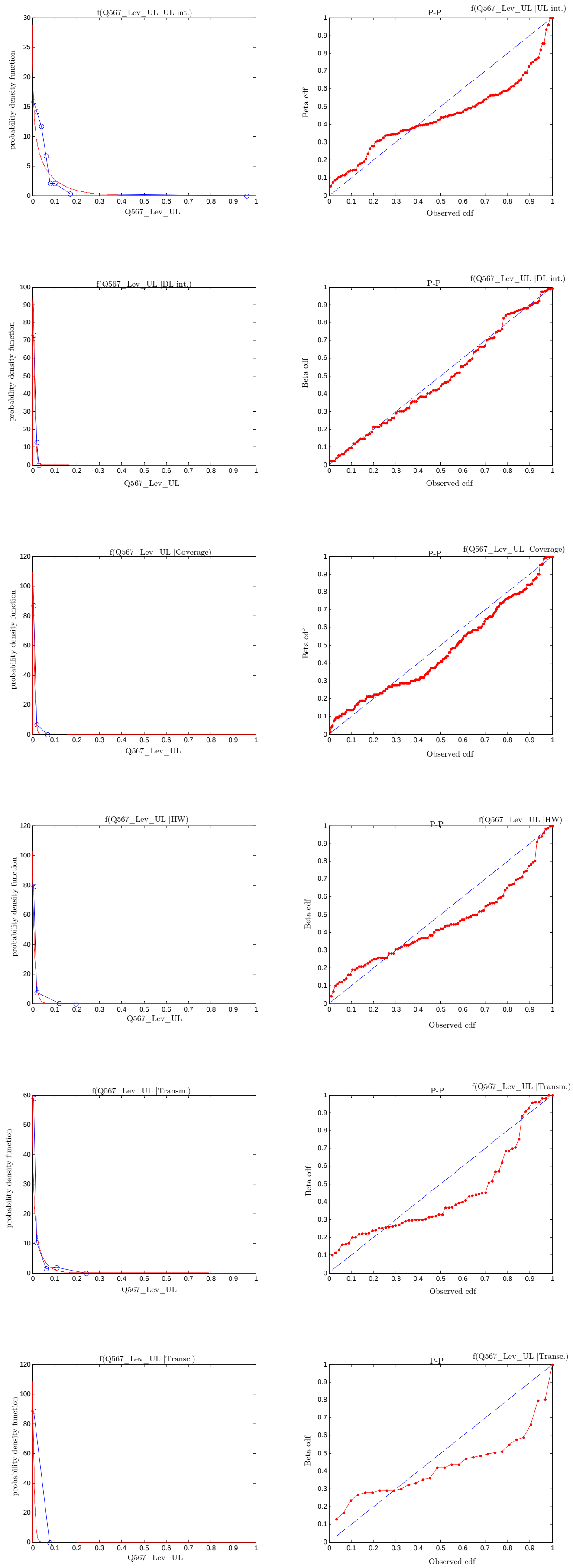


Figure B.14: Modelling of pdfs for symptom "Q567_Lev_UL"

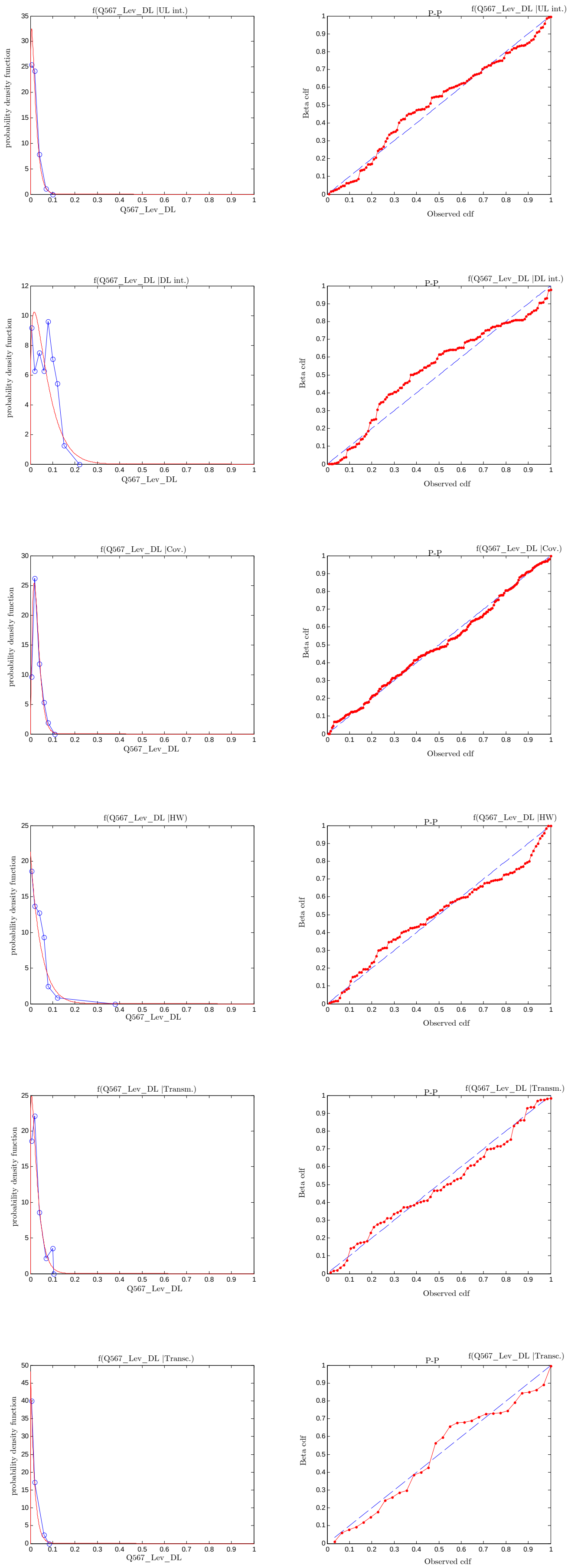


Figure B.15: Modelling of pdfs for symptom "Q567_Lev_DL"

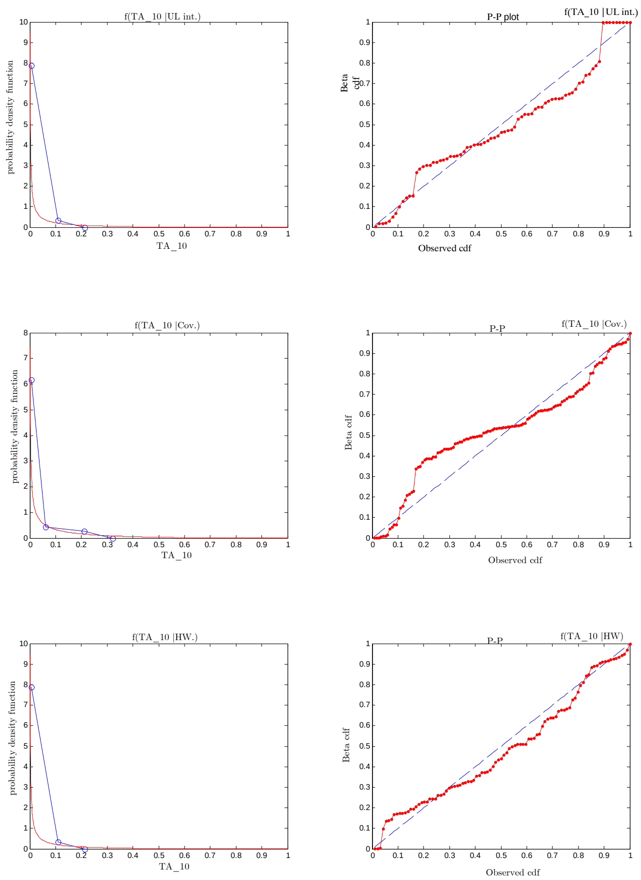


Figure B.16: Modelling of pdfs for symptom “TA_10”

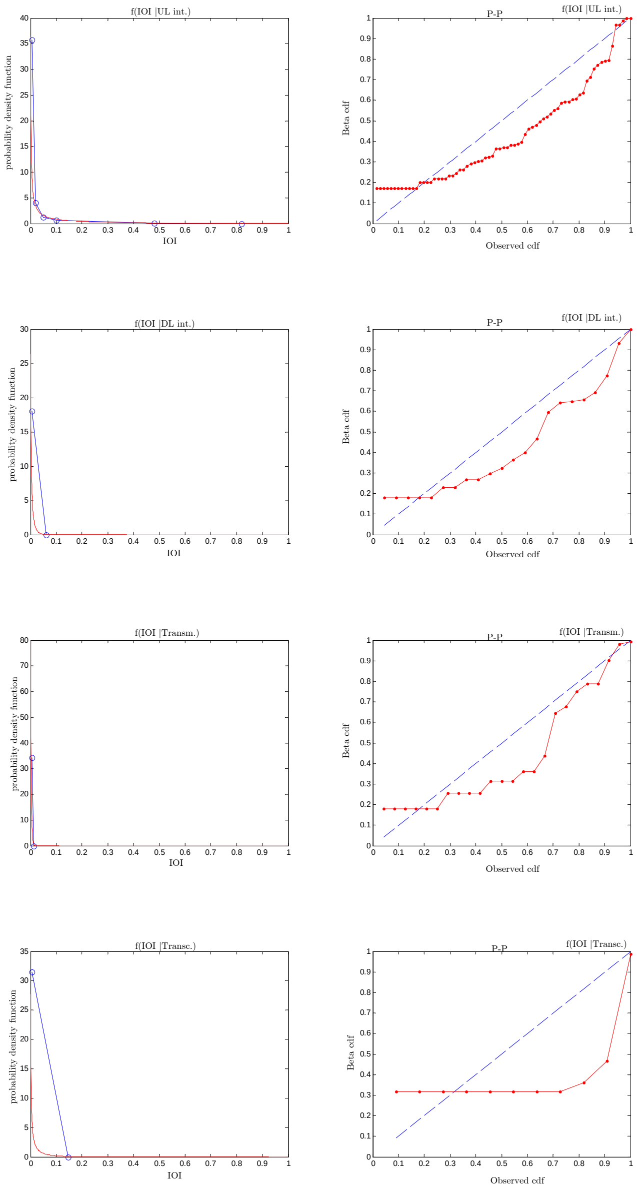


Figure B.17: Modelling of pdfs for symptom “IOI”

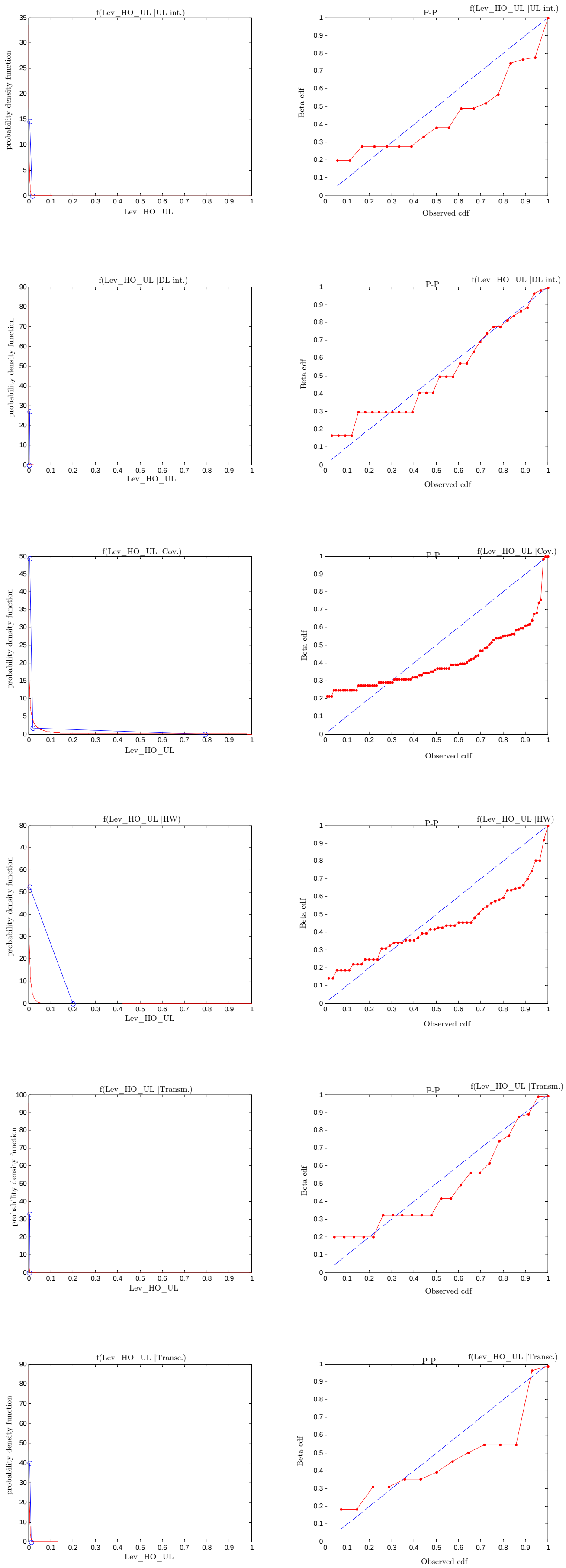


Figure B.18: Modelling of pdfs for symptom “Lev_HO_UL”

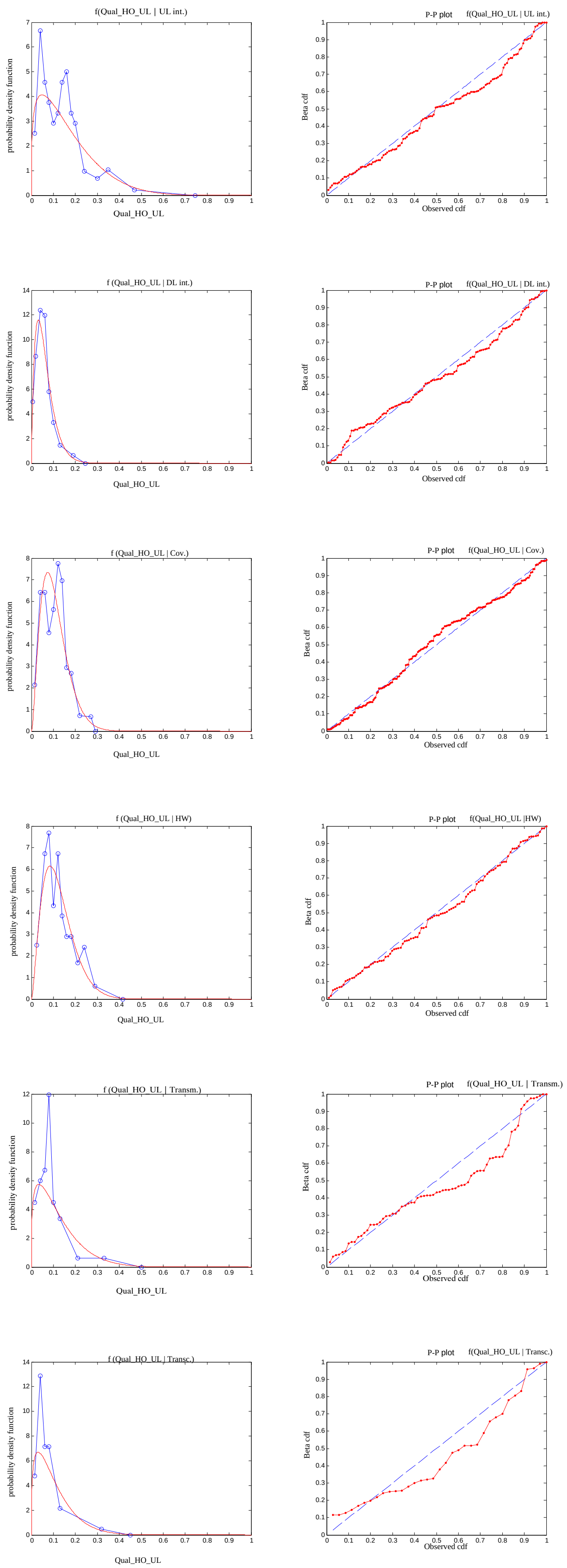


Figure B.19: Modelling of pdfs for symptom "Qua_HO_UL"

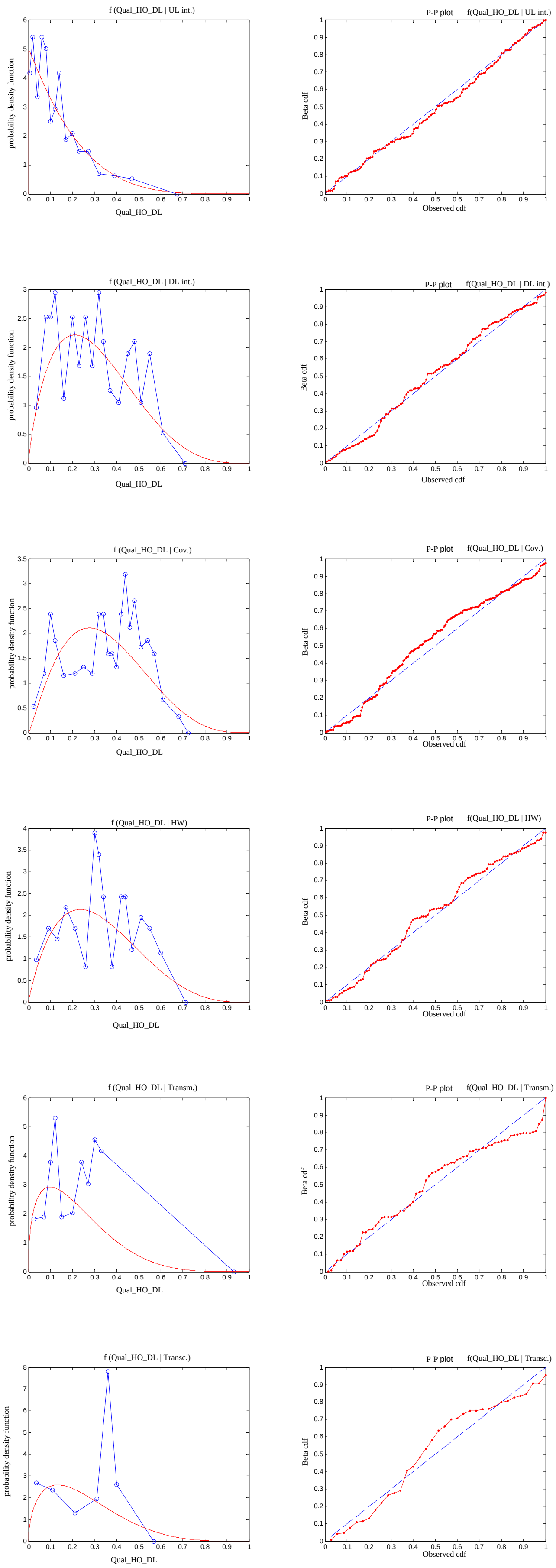


Figure B.20: Modelling of pdfs for symptom "Qua_HO_DL"

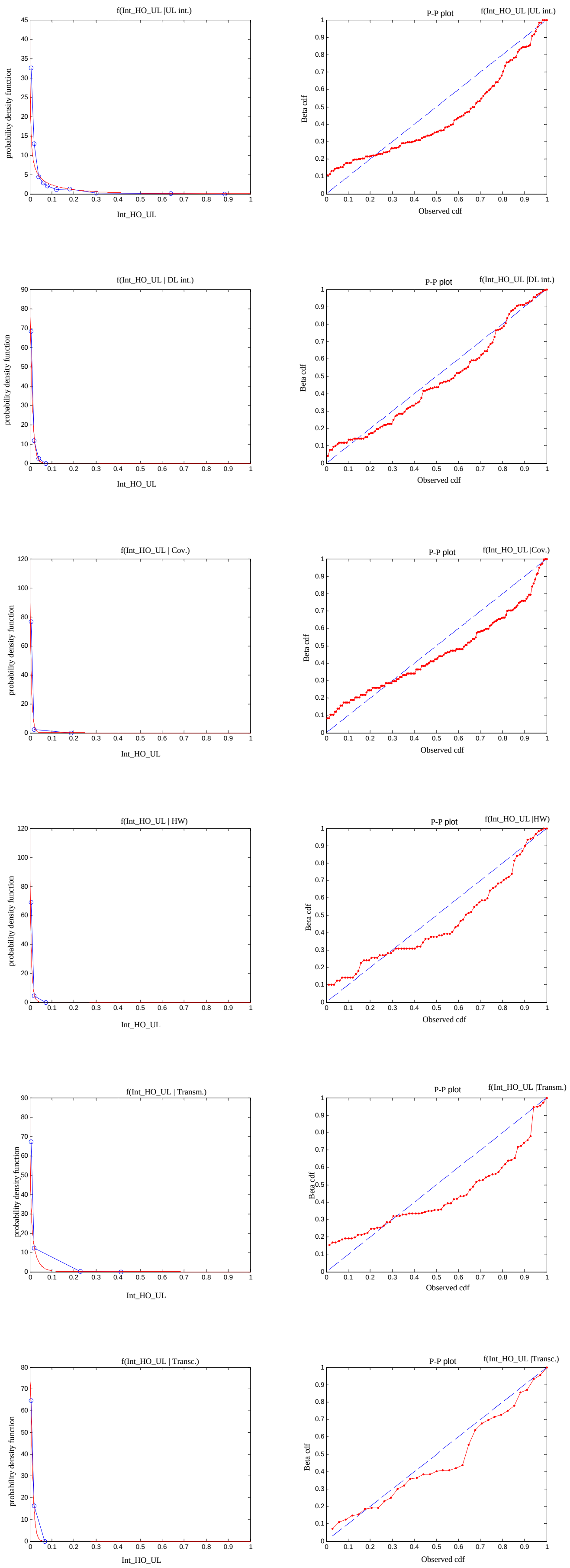


Figure B.21: Modelling of pdfs for symptom “Int_HO_UL”

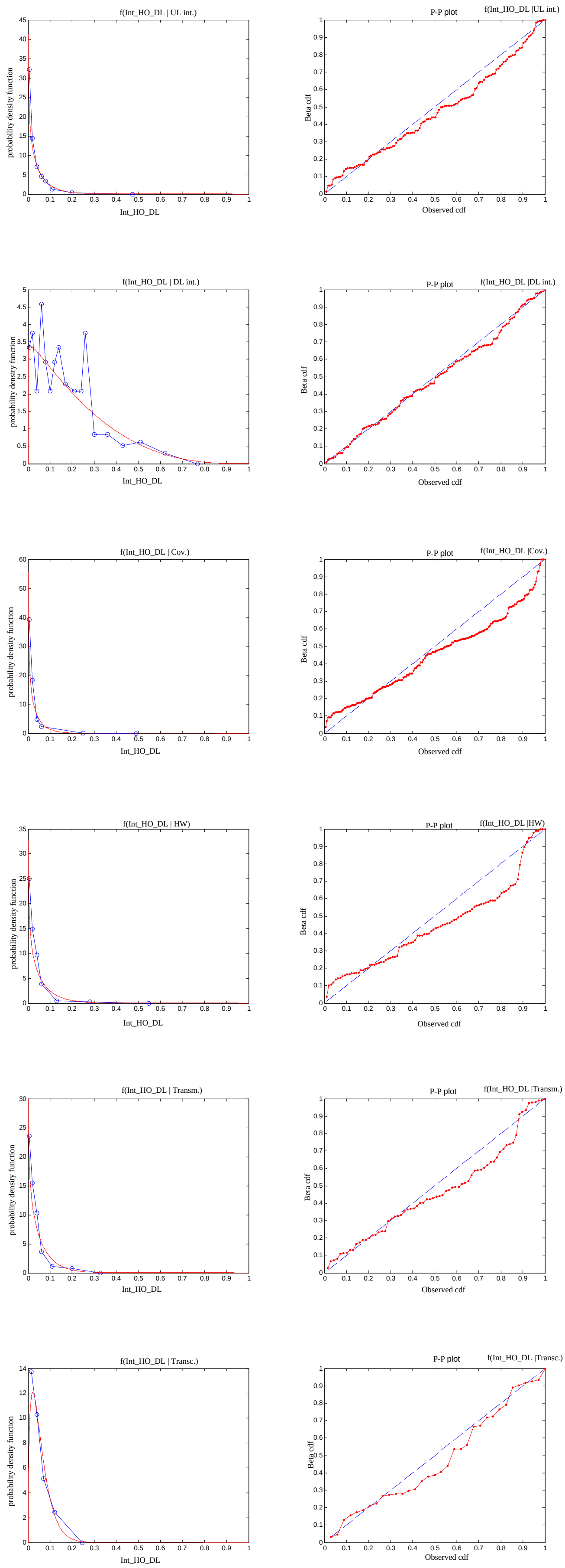


Figure B.22: Modelling of pdfs for symptom “Int_HO_DL”

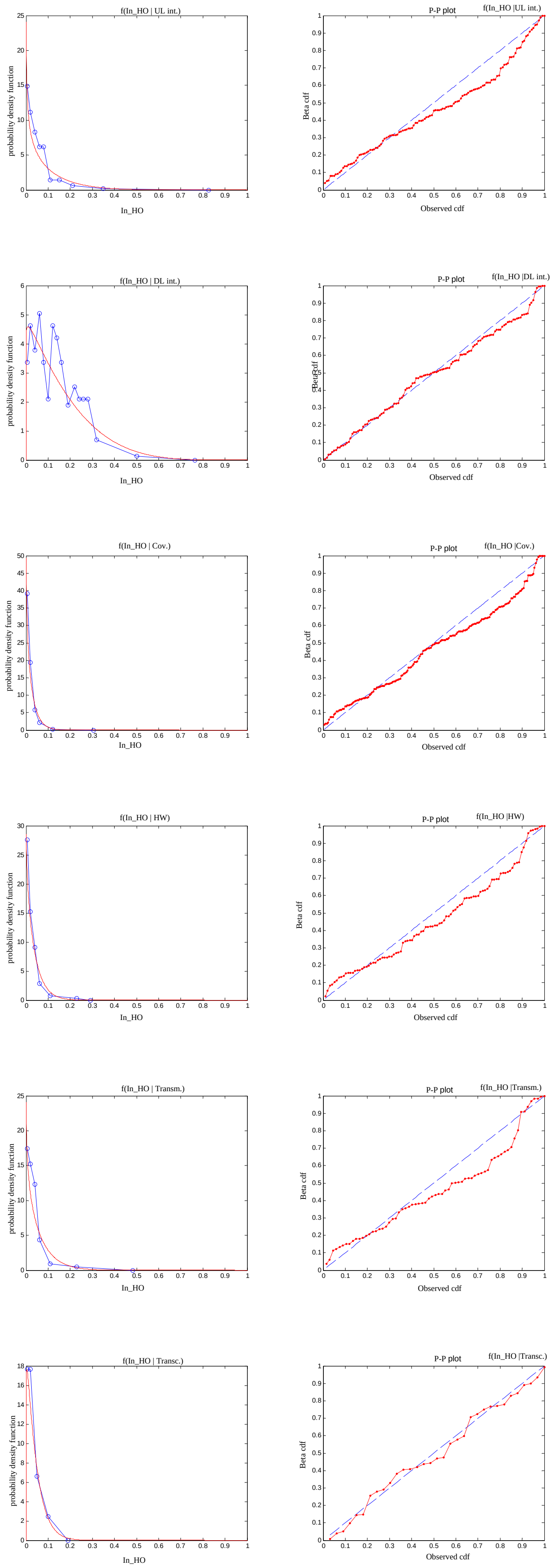


Figure B.23: Modelling of pdfs for symptom “In_HO”

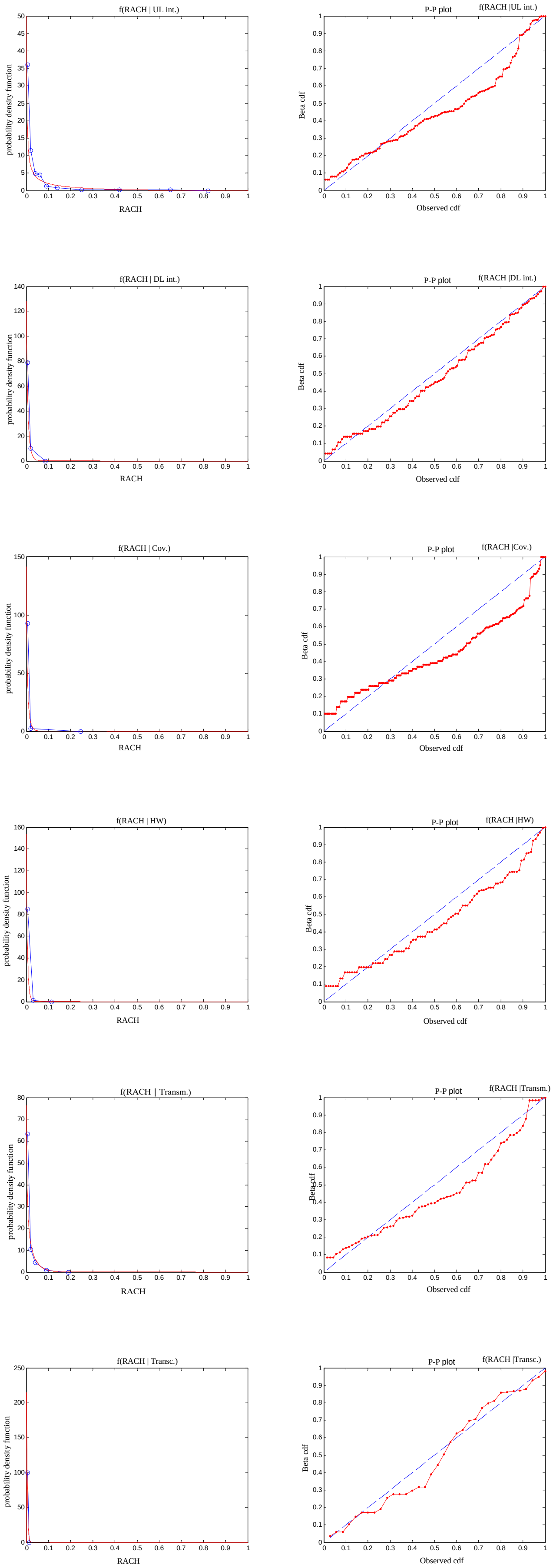


Figure B.24: Modelling of pdfs for symptom “RACH”

Appendix C

Reference pdfs of the symptoms given the causes

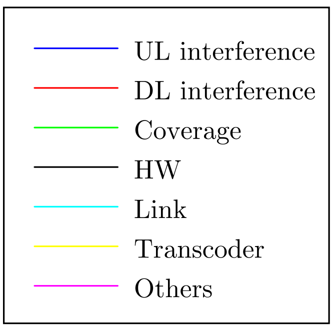


Figure C.1: Legend for figures below

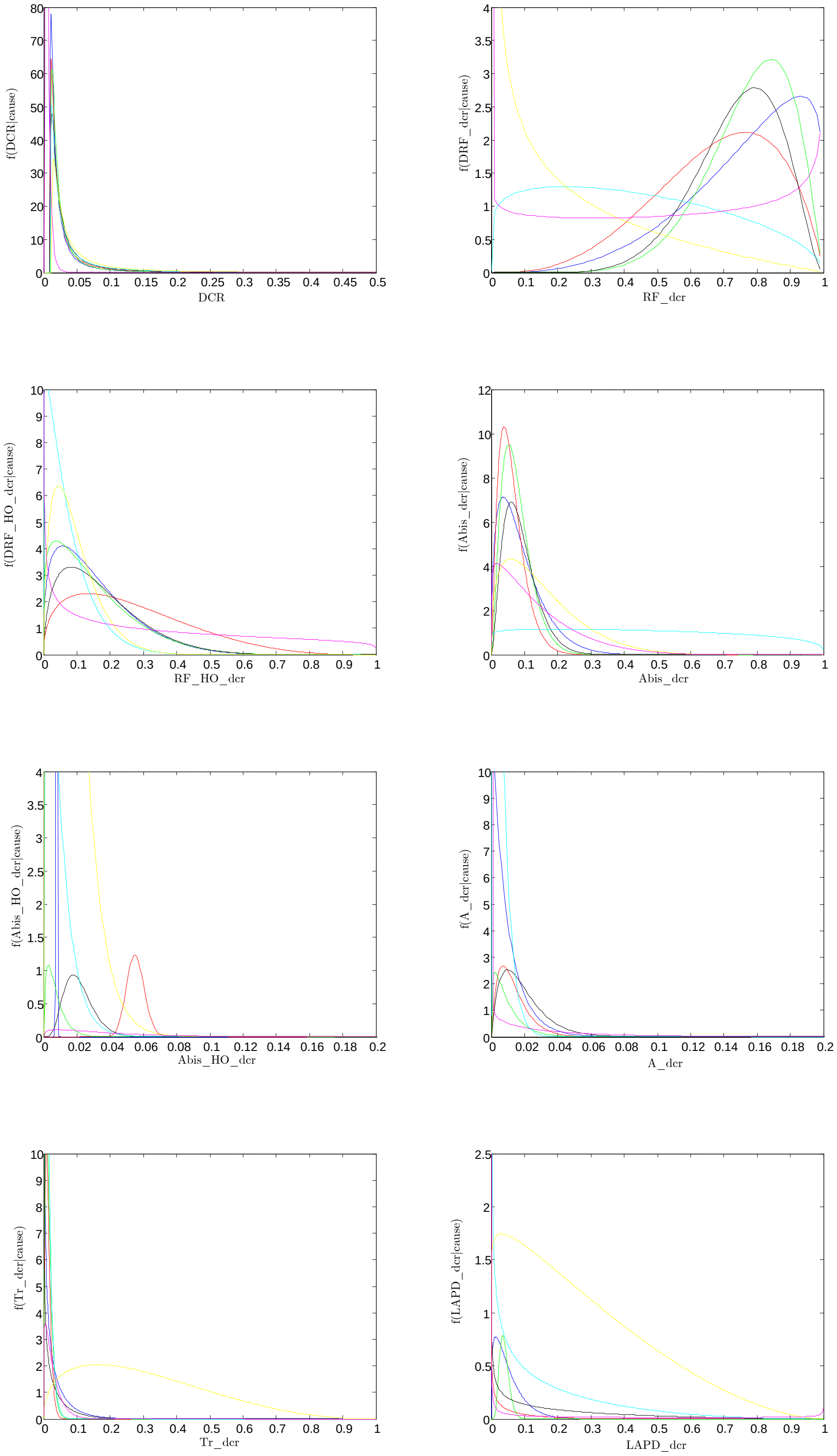


Figure C.2: Pdf of symptoms (I)

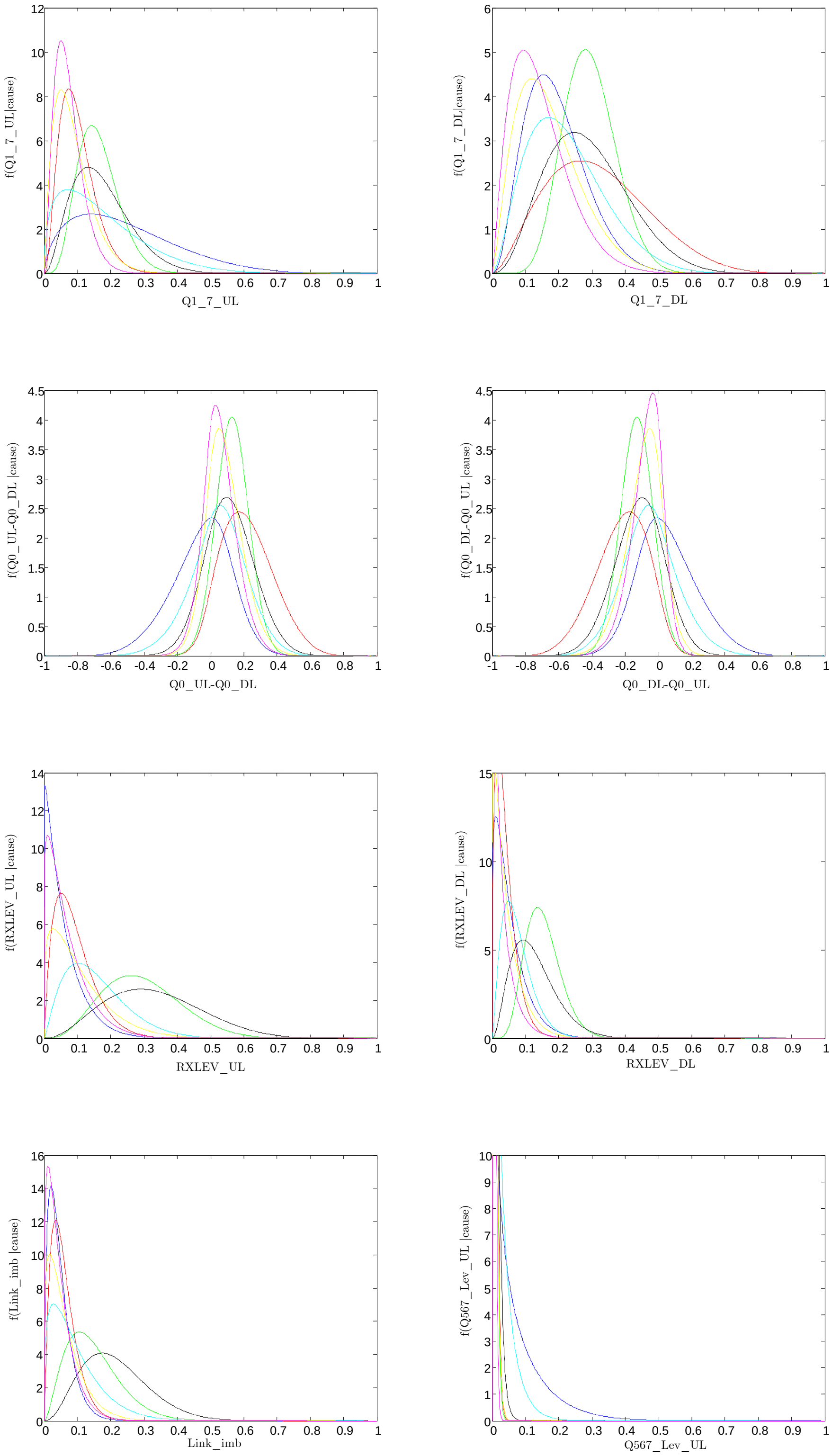


Figure C.3: Pdf of symptoms (II)

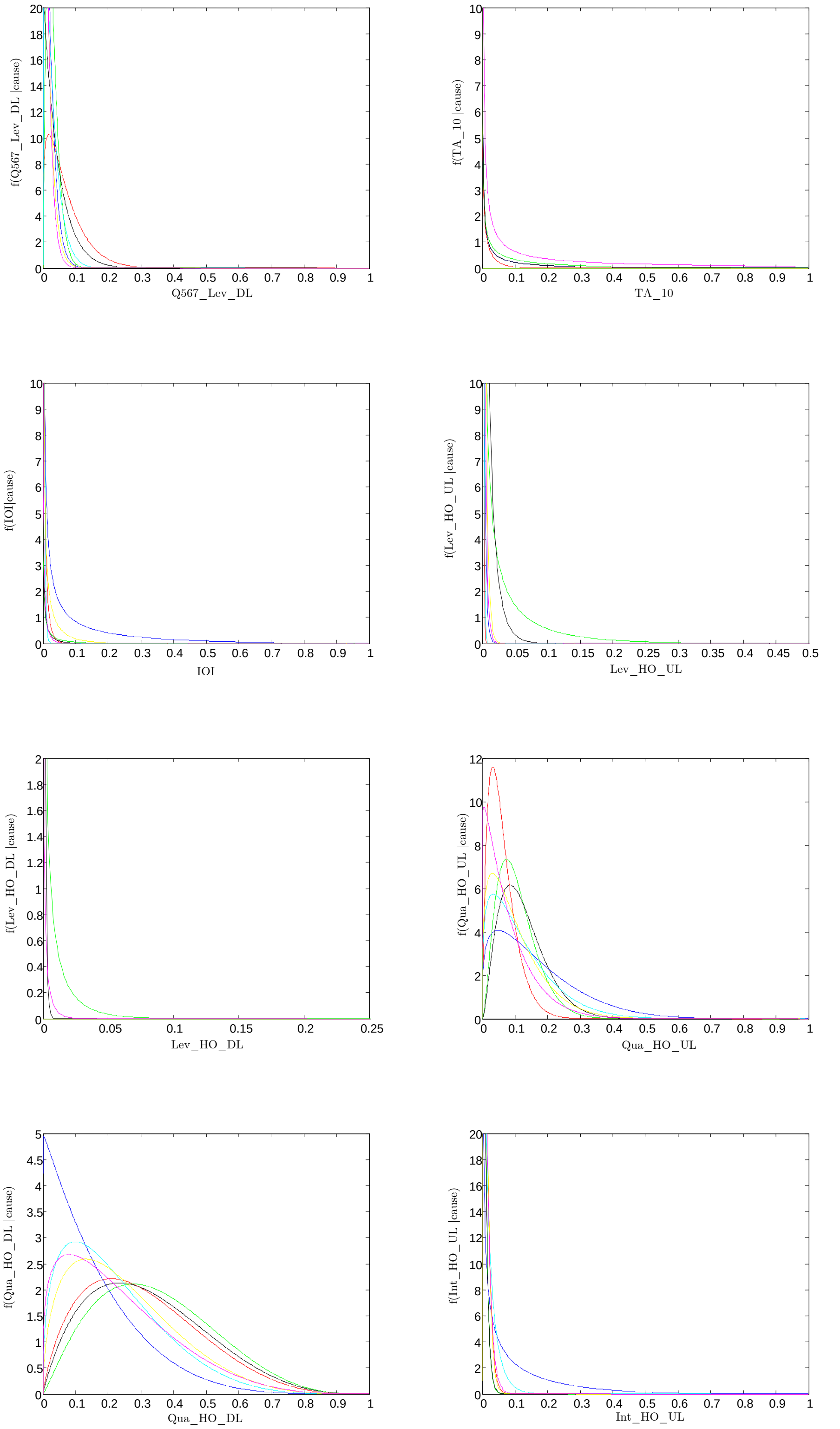


Figure C.4: Pdf of symptoms (III)

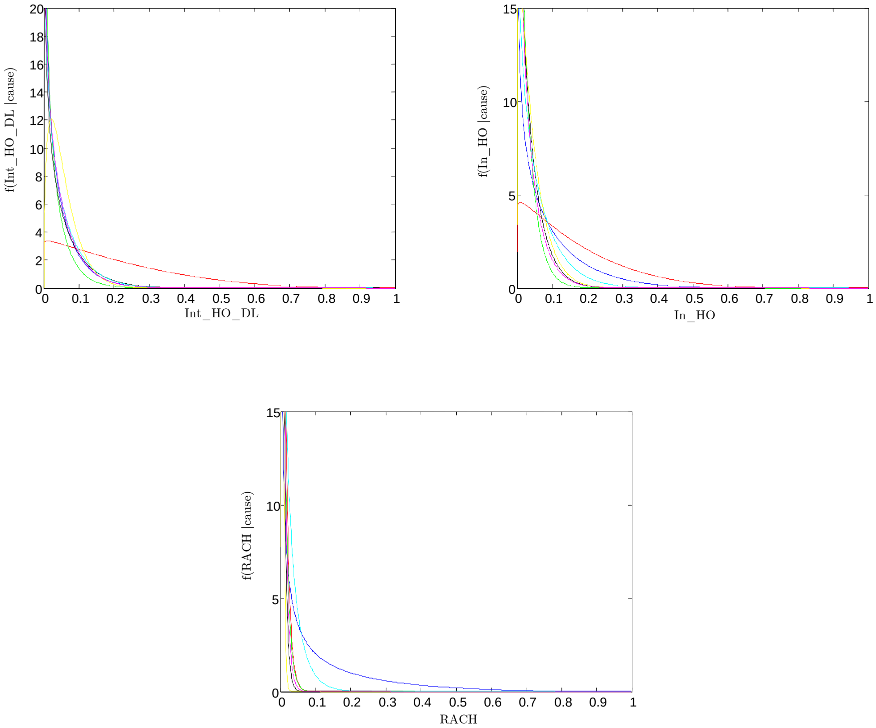


Figure C.5: Pdf of symptoms (IV)

MENÚ SALIR

Appendix D

Diagnosis model defined for field test

Table D.1: Causes in diagnosis model

Ner	Cause
1	Interference in UL
2	Interference in DL
3	Interference in UL in TRX n (BB hopping)
4	Interference in UL in TRX n (no hopping)
5	Interference in DL in TRX n (BB hopping)
6	Interference in DL in TRX n (no hopping)
7	HW fault (combiner or TRX) (Hopping)
8	HW fault (combiner or TRX) (Non hopping)
9	Bad coverage (borders)
10	Bad coverage (possible hole: buildings, indoors, etc.)
11	Gross antenna misalignment or crossed feeders
12	HW problem (feeders, connectors, etc.)
13	Bad target cell coverage
14	HW problem in target cell
15	Problem in target cell: interference, blocking, etc.
16	Gross configuration problem in HO parameters
17	Additional adjacencies needed
18	Problem in the A interface
19	HW fault (transcoder)
20	Other HW fault causing transcoder failures
21	LAPD failure
22	BSCU resets causing increased DCR
23	Abis problem
24	Gross incorrect cell parameter setting
25	TRX fault
26	HW fault (other)
27	HW fault (antenna)
28	RX path fault (RX fail, MHA fault, etc.)
29	Other fault

Table D.2: Conditions in diagnosis model

Conditions
Cell density
Cell Type
Frequency Hopping
Target cell density
Weather

Table D.3: Symptoms in diagnosis model

Symptoms
Increased DCR
Increased radio failure in old channel in HO component in DCR
Increased radio failure component in DCR
Increased A interface failure in DCR
Increased A interface old channel failure in DCR
Increased Abis interface failure in DCR
Increased Abis interface old channel failure in DCR
Increased LAPD failure in DCR
Increased transcoder failure in DCR
Increased number of HO failures
Increased number of UL or DL level HOs
Increased number of UL or DL quality HOs
Increased number of UL or DL interference HOs
Increased number of DL quality HOs
Increased number of UL quality HOs
Increased number of DL interference HOs
Increased number of UL interference HOs
Increased number of intracell HOs
IOI levels increased
Rx quality Q567 samples UL or DL $>$ 6% single TRX
Rx quality 5 XOR 6 XOR 7 UL $>$ 5% single TRX
UL or DL BER in band 0 $<$ 60%
UL Q0 BER samples $<$ DL Q0 BER samples
Reduction in number of samples compared to other TRXs on sector
$>$ 35% of calls out of BER band 0 DL
$>$ 35% of calls out of BER band 0 UL
Blocking
Averg.serving distance of single TRX $<$ 50% averg.value for the other TRXs on sector
UL signal level $<$ DL signal level on all TRXs by more than 10 dBs
Average DL signal level reduced
Average UL signal level reduced
Averg.DL signal level of single TRX is 10 dB $<$ averg.other TRXs on sector
UL signal strength $>$ DL signal strength
Reduction in traffic 24h
$>$ 15% samples $>$5.5 Km from TA figures
Low average DL signal level when low average TA
Calls not being handled by expected adjacencies
Other cells in the same link are affected (A interface)
Other cells in the same link are affected (Abis interface)
No other cells in the same BSC are affected (Transcoder)
Excessive interference alarms
Abis alarms
Antenna combiner alarms
Mains fail alarms
Output power decreased alarm
Channel failure rate alarm occurrence
LAPD alarms
BSCU alarms
Transcoder alarms
HW alarms
A interface alarms

Appendix E

Summary (Spanish)

E.1 Introducción

E.1.1 Antecedentes y Justificación

En el año 2000, Nokia Networks inauguró un Centro de Ingeniería de Sistemas de Comunicaciones Móviles en el Parque Tecnológico de Andalucía (PTA). El proyecto se llevó a cabo gracias a un acuerdo de colaboración entre la Universidad de Málaga, en particular el Departamento de Ingeniería de Comunicaciones, y Nokia. El personal del Centro estuvo formado por trabajadores de Nokia procedentes principalmente de Finlandia e Inglaterra, más de 50 personas de nueva contratación y profesores de la Universidad de Málaga.

Como uno de los profesores que trabajó a tiempo parcial en Nokia, me integré en un nuevo grupo de trabajo formado por otras tres personas en Málaga, una en Inglaterra y una en Dinamarca. El objetivo del grupo era diseñar un sistema automatizado de resolución de problemas para la Red de Acceso Radio (RAN) de sistemas de comunicaciones móviles. Por los motivos que se expondrán en E.1.2, Nokia consideró prioritario abrir una línea en este campo de investigación.

Debido a que previamente trabajé en diagnosis de estaciones terrenas de seguimiento de satélites [44, 41, 131] en la Agencia Espacial Europea (ESA), me pareció muy interesante incorporarme a esta línea de investigación. Por otra parte, la tarea era complicada y prometedora al mismo tiempo, puesto que no existían estudios previos relacionados con la diagnosis en redes de acceso de sistemas celulares. Además, se trataba de un tema multidisciplinar para el que no sólo se requerían conocimientos de comunicaciones móviles, sino también de inteligencia artificial.

Mi labor se desarrolló en las propias instalaciones de Nokia en el PTA, siendo la responsable de las tareas de investigación del grupo. Las funciones de los demás integrantes del equipo incluían la gestión del proyecto, el desarrollo de aplicaciones y las relaciones con operadores de redes de comunicaciones móviles.

La experiencia fue muy gratificante, no solamente me dediqué a la investigación, sino que estuve directamente implicada en la definición de prototipos de herramientas para la diagnosis y en los contactos con operadores de telefonía (realizando diversas reuniones con operadores en Inglaterra y Dinamarca). Lamentablemente, en Junio de 2003 Nokia tuvo que tomar la decisión política de cerrar el centro de Málaga, con lo cual el proyecto se disolvió.

No obstante, la necesidad de herramientas automáticas de diagnosis seguía existiendo y los objetivos de mi tesis empezaban a estar claros. Por este motivo seguí investigando en el mismo tema.

En Septiembre del 2003 el Ministerio de Ciencia y Tecnología aprobó un proyecto de investigación dentro del Grupo Ingeniería de Comunicaciones denominado "Desarrollo de herramientas de optimización de los recursos radio en redes de comunicaciones móviles", una de cuyas líneas es la diagnosis en la RAN.

Además, desde el año 2004 al 2007 he participado en un consorcio, entre los que han formado parte, entre otros, France Telecom y Telefónica I+D, para llevar a cabo un proyecto sobre monitorización y autorregulación de parámetros de gestión de recursos radio (RRM) en una red multi-sistema, que obtuvo la etiqueta CELTIC [1] dentro de la red europea EUREKA [4]. Este proyecto ha incluido una tarea relacionada con la diagnosis en la RAN, de la que la Universidad de Málaga es responsable.

E.1.2 Formulación del problema y objetivos

La industria de telecomunicaciones móviles está experimentando en los últimos tiempos cambios extraordinarios gracias a la introducción de nuevas tecnologías y servicios, y a los altos niveles de competencia. En los próximos años nuevas funcionalidades deberán integrarse en las redes actuales para permitir la convergencia de los sistemas celulares con Internet. En este escenario de complejidad creciente, los operadores y fabricantes de equipos para comunicaciones móviles están haciendo enormes esfuerzos para adaptar las redes celulares a las nuevas tecnologías y, al mismo tiempo, mantener el nivel de servicio de las redes actuales. En consecuencia, la operación de la RAN es cada vez más complicada.

Además, se vislumbra que en un futuro próximo nuevas tecnologías van a revolucionar la industria de telecomunicaciones. Por ejemplo, la expansión creciente de la telefonía sobre IP, la extensa disponibilidad de acceso WLAN y el lanzamiento de la telefonía basada en dichas tecnologías, está suponiendo una amenaza potencial para el negocio de los operadores de comunicaciones móviles. Por tanto, estas empresas deben buscar formas de reducir sus gastos y mejorar su calidad para hacer frente a la amenaza de las tecnologías emergentes.

En el pasado, los operadores consiguieron hacer frente a los rápidos cambios tecnológicos incrementando su personal. Sin embargo, debido a las presiones financieras, esta estrategia ya no es factible y la única opción viable para reducir los costes operacionales es aumentar el grado de automatización. Por eso, actualmente dichos operadores están mostrando un enorme interés por automatizar la gestión de la red, con el objetivo de incrementar la eficiencia operacional. La automatización se puede aplicar en áreas muy diferentes de la RAN de los sistemas celulares [195, 186, 38, 194, 185, 125, 134, 135, 113, 177]: planificación de frecuencias, definición de adyacencias, resolución de problemas, optimización de parámetros, etc. Uno de los primeros pasos en la automatización de la red es la resolución de los problemas, ya que si una celda temporalmente está fuera de servicio, el funcionamiento de las celdas vecinas probablemente también se vea afectado, dando como resultado muchos clientes insatisfechos.

Hasta el momento la diagnosis de fallos ha sido una tarea manual, que requiere personal dedicado de forma exclusiva al análisis de cientos de indicadores de funcionamiento de la red y alarmas. Estos "expertos" en resolución de fallos se encargan de diagnosticar cuál es la causa de los problemas que ocurren en ciertas celdas, de forma que puedan solucionarse lo antes posible.

El **objetivo general** de esta tesis es el diseño de un sistema automático de diagnosis para redes de acceso de sistemas celulares. Los **objetivos específicos** son los siguientes:

- Aproximación al problema. Como primera etapa, se debe llevar a cabo una investigación sobre la forma de realizar la gestión de fallos en las redes actuales. Además, se deberán examinar diversas técnicas dentro del campo de la inteligencia artificial para la diagnosis automática, que tengan en cuenta la incertidumbre inherente al razonamiento humano.

- Diagnosis en redes celulares. Se debe proponer un sistema de diagnosis automático para las actuales redes de comunicaciones móviles (GSM/GPRS). Para ello, se deben identificar las principales variables a tener en cuenta y las relaciones entre ellas.

- Modelado de la diagnosis. Se deben proponer métodos para diagnosticar de forma automática la causa de los problemas en cualquier red celular. Dichos métodos utilizan un *modelo* de diagnosis que representa el conocimiento sobre cómo averiguar la causa de los problemas. Como parte de este objetivo, se deben identificar los principales elementos del modelo y sus parámetros.

- Construcción del modelo. Para definir los parámetros del modelo hay dos soluciones. Por una parte, el modelo puede ser definido por expertos en la RAN. Por tanto, un objetivo es estudiar cómo convertir las especificaciones en lenguaje natural proporcionadas por los expertos en modelos de diagnosis. Esto se conoce como *adquisición del conocimiento*. Por otra parte, el modelo se puede aprender a partir de ejemplos de entrenamiento. Por eso, otro objetivo es desarrollar métodos para aprender automáticamente los parámetros a partir de dichos casos de entrenamiento.

- Evaluación del modelo. El último objetivo es diseñar técnicas para evaluar y comparar los sistemas de diagnosis.

Aunque en esta tesis los métodos que se proponen se aplican a redes GSM, ya que de ellas se pueden obtener datos reales hoy en día, estas técnicas son generales y, por tanto, pueden aplicarse también a otros sistemas móviles, como UMTS.

El documento de la tesis se ha dividido en tres partes:

- Elementos del marco teórico. Esta fase exploratoria incluye dos aspectos principales: revisión de literatura relacionada y resumen de los fundamentos teóricos. A su vez, debido al carácter multidisciplinar de la tesis, esta parte se ha dividido en dos capítulos claramente diferenciadas dentro de las disciplinas de comunicaciones móviles e inteligencia artificial. En primer lugar es fundamental el estudio del estado actual de la automatización de redes de comunicaciones móviles. En particular, se estudia cómo se realiza actualmente la diagnosis de fallos en redes de acceso GSM. El segundo capítulo se centra en el estudio de técnicas usadas para la diagnosis automática en otros campos, con el objetivo de seleccionar la(s) que se considere más apropiada(s) para los sistemas celulares.

- Una vez conocido el estado del arte, se pasa al diseño propiamente dicho del sistema de diagnosis. La segunda parte de la tesis comprende también dos capítulos: diagnosis en redes GSM y técnicas para la diagnosis automática. El primer capítulo consiste en investigar las principales causas de problemas, sus síntomas y otros factores relacionados en GERAN. El segundo capítulo propone modelos y métodos, aplicables a cualquier sistema celular, para llevar a cabo la diagnosis automática.

- La tercera parte de la tesis se dedica a la evaluación y comparación de los modelos y métodos propuestos en la segunda parte. Se expone también la metodología seguida para la obtención de datos y su análisis. Finalmente se extraen conclusiones y se proponen líneas de acción futuras.

E.2 Estado del arte

La primera parte de la tesis presenta el marco teórico en el que se sitúa la tesis. Sus objetivos son los siguientes:

- Presentar los fundamentos teóricos sobre comunicaciones móviles y sistemas expertos necesarios para comprender el resto de la tesis.
- Revisar la literatura sobre automatización en redes de comunicaciones móviles y diagnosis automática en otros campos de conocimiento.
- Presentar los conceptos más importantes que se utilizarán a lo largo de la investigación.

E.2.1 Automatización en redes de comunicaciones móviles

En la primera parte del Capítulo 2 se resumen los principios de los sistemas GSM, prestando especial atención a aquellos aspectos de GSM requeridos para seguir el resto de la tesis [142, 161, 162, 145, 90, 7].

A continuación, se resalta la importancia de la automatización de las tareas de gestión de red como la mejor forma de reducir costes operacionales en las actuales redes de comunicaciones móviles de creciente complejidad [195, 134, 135, 113, 177, 90, 125].

Se describe en qué consiste la resolución de problemas (*troubleshooting, TS*) en redes de acceso de sistemas de comunicaciones móviles. Para ello, se explican cada una de las fases de que consta el TS: identificación de celdas con problemas, diagnosis, acciones para la resolución del problema. Se describe cómo la mayor parte de los operadores realizan en la actualidad el TS ayudándose de un sistema de gestión de incidencias (*troubleticket*) [136, 139, 138], que permite describir los fallos que aparecen en la red y asignar su resolución a la persona o grupo más adecuado para tratarlo. Asimismo se describe cómo se realiza la diagnosis de fallos de forma manual, con la ayuda de herramientas de visualización que permiten el análisis de los principales indicadores de funcionamiento y alarmas.

Una vez justificada la necesidad de una herramienta automática de diagnosis y comprendido cómo se lleva a cabo la diagnosis de forma manual, se describe en qué consistiría la diagnosis automática en redes celulares y el escenario en el que se localizaría dicha herramienta automática. Posteriormente, se estudian los primeros pasos llevados a cabo en la automatización de la resolución de problemas: la detección automática de fallos [123, 124, 104, 133, 48, 47] y la correlación de alarmas [109, 202, 192, 84, 88, 193]. Aunque la correlación de alarmas (interpretación de múltiples alarmas, de forma que se asigne un nuevo significado a esas alarmas) puede entenderse como un primer paso en la diagnosis de fallos, normalmente las alarmas no proporcionan información concluyente para identificar la causa de los problemas, especialmente si los problemas no son sólo fallos en equipos. Otras categorías de fallos, como interferencia o falta de cobertura, son difíciles de identificar si no se consideran también los indicadores de funcionamiento. Finalmente se revisan los estudios llevados a cabo sobre diagnosis automática en otras áreas de conocimiento, como la medicina [34, 146, 99, 199, 35, 154, 147, 144] o el núcleo de redes de comunicaciones [115, 182, 190, 73].

E.2.2 Técnicas para la diagnosis automática

En el Capítulo 3, se realiza un estudio sobre técnicas usadas para la diagnosis automática en otras áreas de conocimiento, en particular, se investigarán aquellos sistemas basados en el conocimiento capaces de modelar la incertidumbre (teoría de Dempster-Shafer [72, 169], lógica borrosa [205, 65], factores de certidumbre [172, 52, 92], redes bayesianas [157, 110, 54, 64], etc.). De entre todas estas técnicas se justifica la elección de la aproximación bayesiana adoptada en esta tesis. En la segunda parte del capítulo, se presenta una breve introducción a las Redes Bayesianas.

E.3 Modelado de la diagnosis de fallos

En la segunda parte de la tesis se presentan sus principales aportaciones. El objetivo es diseñar un sistema de diagnosis para la RAN de sistemas de comunicaciones móviles, el cual se compone de dos partes fundamentales: un modelo y un método de inferencia. El *modelo* representa el conocimiento del experto sobre el dominio de aplicación, en este caso el conocimiento de expertos sobre cómo se lleva a cabo la identificación de la causa de los problemas en la RAN. Hay dos aspectos involucrados en la construcción del modelo: la información sobre el dominio de aplicación (*base de conocimiento*) y su *representación*. El *método* es el algoritmo de inferencia que identifica la causa de los problemas basándose en la evidencia disponible.

Aunque la base de conocimiento estudiada es la relacionada con la diagnosis en redes GSM/GPRS, los modelos y métodos que se proponen son también válidos para otros sistemas celulares con sólo cambiar los elementos del modelo.

El Capítulo 4 estudia la base de conocimiento sobre diagnosis en redes GSM/GPRS, es decir, se analiza en qué consiste la diagnosis en este tipo de redes. En el Capítulo 5 se investigan representaciones del modelo y métodos bayesianos para la construcción del sistema de diagnosis.

E.3.1 Diagnosis en redes GSM/GPRS

Introducción

En primer lugar se definen los principales conceptos que se utilizarán en este capítulo (problema, fallo, causa, síntoma, etc.).

La tesis se centra en la diagnosis de las causas cuando el problema es un elevado número de llamadas caídas. La tasa de llamadas caídas es uno de los principales indicadores de funcionamiento usados por los operadores, ya que una llamada caída puede tener un impacto muy negativo en el servicio final ofrecido al cliente. Se estudian distintas figuras de mérito usados por los operadores para medir las llamadas caídas.

Causas

A continuación se analizan las principales causas, relacionadas con la RAN, que en un sistema GSM/GPRS provocan una tasa elevada de llamadas caídas: falta de cobertura, interferencia, fallos hardware, fallos de transmisión, etc.

Síntomas

En esta sección se estudiarán los síntomas considerados para la diagnosis, tanto indicadores de funcionamiento como alarmas. El Sistema de Gestión de Red (*Network Management System*, NMS) almacena diariamente los indicadores de funcionamiento más importantes, gracias a contadores situados en diversos puntos de la red. Además, el NMS proporciona información sobre cientos de alarmas de los elementos de la red. Cuando una de las causas descritas en la sección anterior provoca una tasa elevada de llamadas caídas en una celda, los valores de algunos indicadores de funcionamiento cambian respecto a sus valores nominales y diversas alarmas se disparan. Los principales indicadores de funcionamiento se relacionan con la calidad y nivel de la señal recibida, causas de traspasos, etc.

Condiciones

Por último, se describen las condiciones, que son factores que pueden influir en la aparición de un determinado fallo. Las condiciones pueden agruparse en funcionalidades y configuraciones. Las *funcionalidades* son técnicas que los operadores normalmente aplican para mejorar la calidad de la red, como los saltos de frecuencia, la transmisión discontinua, etc. Las *configuraciones* son características especiales de ciertas celdas que pueden tener impacto sobre las causas que normalmente se presentan, por ejemplo, el tipo de celda (rural/urbana, interior/exterior, etc.), el clima, etc.

Relación cualitativa entre causas, síntomas y condiciones

Una vez que se han explicado las principales causas, síntomas y condiciones a tener en cuenta en la GERAN, se describe la relación cualitativa que existe entre estas variables. Por ejemplo, la causa "falta de cobertura" va asociada a los síntomas "calidad de señal recibida" y "nivel de señal recibida", ya que su presencia provoca una disminución del valor de estos síntomas.

Casos de estudio

Se presentan algunos casos de redes reales. En primer lugar, se muestran los valores de los indicadores de funcionamiento para celdas con problemas. A continuación, se describen algunos ejemplos de celdas con problemas cuya identificación es especialmente compleja, y su resolución por expertos.

E.3.2 Modelado bayesiano de la diagnosis de fallos

Introducción

Este capítulo se centra en la representación del modelo y en el método de inferencia. En primer lugar, se describen las variables que deben incluir todos los modelos (causas, síntomas y condiciones). Se explica cómo modelar estos elementos mediante variables aleatorias, algunas continuas y otras discretas. Se resalta que una de las mayores dificultades encontradas en el modelado se debe a que los indicadores de funcionamiento son variables continuas.

Clasificador Bayesiano

El primer sistema que se propone consiste en un *clasificador bayesiano*, el cual aparece con frecuencia en referencias sobre aprendizaje de máquinas (*machine learning*) [165, 74, 82]. El clasificador bayesiano es uno de los clasificadores más eficientes, ya que, a pesar de ser muy simple, es competitivo con el resto de los clasificadores actuales.

Se estudia cómo utilizar el clasificador para la diagnosis, sus limitaciones y los elementos de los que consta el modelo. La parte cualitativa del modelo se define mediante variables aleatorias. La parte cuantitativa queda determinada por funciones densidad de probabilidad (fdp).

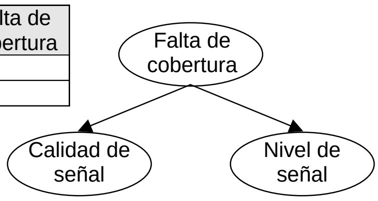
Se investiga cómo obtener las fdp en el caso de la diagnosis en la RAN, demostrando que las fdp de los síntomas dadas las causas pueden modelarse de forma bastante precisa utilizando fdp beta. Los parámetros de las fdp pueden obtenerse a partir de casos de entrenamiento o bien ser proporcionados por expertos en diagnosis.

Por último, se presenta el método para calcular las probabilidades de las causas dados los síntomas, el cual se basa en la conocida regla de Bayes.

Redes Bayesianas

Las *Redes Bayesianas* (RBs) [110, 157], también llamadas redes de creencia probabilísticas (*belief probabilistic networks*), han sido propuestas por muchos autores como la técnica de modelado

Falta de cobertura
No
Sí



	Falta de cobertura	
Calidad de Señal	No	Sí
Alta	0.95	0.3
Baja	0.05	0.7

	Falta de cobertura	
Nivel de Señal	No	Sí
Alta	0.9	0.2
Baja	0.1	0.8

Figure E.1: Ejemplo de Red Bayesiana

para el desarrollo de sistemas de diagnosis. Una RB representa gráficamente una distribución de probabilidad conjunta sobre un conjunto de variables aleatorias. Una de las principales ventajas de las RBs es que proporcionan una representación natural y eficiente del razonamiento bajo incertidumbre.

Los algoritmos de inferencia en RBs (método) permiten obtener la probabilidad de una determinada variable dada la evidencia disponible, basándose en el teorema de Bayes. Existen algoritmos eficientes que realizan esta tarea cuando la RB es compleja.

El modelo consiste en nodos (variables aleatorias), enlaces dirigidos (relaciones de dependencia entre variables) y distribuciones de probabilidad. Por ejemplo, la Fig.E.1 muestra una RB muy sencilla donde se pueden observar los componentes principales de cualquier RB.

Los principales problemas encontrados durante la construcción del modelo han sido la definición de la estructura de la RB, el modelado de los síntomas continuos y la especificación de las probabilidades de la RB:

- Uno de los inconvenientes de las RBs es la dificultad de construir modelos complejos. Por eso, una técnica usada a menudo para simplificar el modelado es asumir una estructura de red, es decir, suponer ciertas relaciones de independencia entre las variables. De esta forma el problema se convierte en la especificación de los nodos y tablas de probabilidad de una determinada RB. Gracias a esto, no sólo se simplifica la definición del modelo, sino también los algoritmos de inferencia.

- Las RBs discretas son más sencillas que las continuas. Por este motivo, son las utilizadas con mayor frecuencia en aplicaciones reales. Por tanto, en el caso bajo estudio es necesaria una discretización de los indicadores de funcionamiento continuos.

- Una vez que las variables continuas se han discretizado, la siguiente dificultad radica en la definición de las tablas de probabilidad de la RB.

- Debido a la falta de bases de datos con casos para aprender los parámetros del modelo, con frecuencia el modelo debe basarse en el conocimiento. Esto da lugar a inevitables imprecisiones en los parámetros. Por eso, se han propuesto dos métodos para disminuir los errores en la diagnosis debido a estas imprecisiones en los parámetros.

Definición de la estructura

En primer lugar, se estudian posibles estructuras de RB adecuadas para la diagnosis en la RAN de redes celulares. Se han elegido por su simplicidad y buenos resultados las siguientes:

- La primera estructura, llamada Modelo Simple de Bayes (*Simple Bayes Model*, SBM), coincide con el clasificador bayesiano. La diferencia es que en este caso los síntomas continuos se han modelado como variables discretas. El SBM consiste en un nodo padre, cuyos estados son las posibles causas, y nodos hijos, que representan los síntomas.

- Otros tipos de RB evaluadas en esta tesis han sido los modelos de Independencia de Influencia Causal (*Independence of Causal Influence*, ICI) [97, 180, 95, 93, 102]. En este caso cada causa se representa como un nodo distinto y las tablas de probabilidad de los síntomas se simplifican bajo suposiciones de independencia causal. Se han considerado distintos modelos ICI, en función de como las causas contribuyen de forma conjunta a los síntomas: Noisy-OR, Noisy-Max, etc.

- Finalmente se propone una estructura, que se ha llamado Modelo de Bayes Central (*Central Bayes Model*, CBM). En este caso, al igual que en el SBM, existe un nodo padre cuyos estados son las posibles causas y nodos hijos que representan los síntomas. Además, cada condición se representa mediante un nodo padre del nodo de causas.

Métodos para la discretización de variables continuas

En este punto se investigan métodos para discretizar los indicadores de funcionamiento. A la hora de seleccionar los métodos se ha tenido en cuenta no sólo la precisión de la clasificación, sino también la complejidad del método.

Se distingue entre métodos basados en la experiencia y métodos basados en casos de entrenamiento. Como aportación a la tesis se proponen dos métodos que combinan la experiencia con la información dada por casos de entrenamiento.

Los métodos propuestos para la RAN son los siguientes:

- El primer método, que se basa únicamente en el conocimiento, es el más adecuado cuando no se dispone de datos de entrenamiento. Los expertos en diagnosis definen los umbrales. Se describen las dificultades asociadas a esta tarea y cómo superarlas.

- Entre los métodos presentes en la bibliografía de aprendizaje de máquinas, destacan los métodos basados en la entropía. Se ha seleccionado un algoritmo que elige como umbral aquel punto que minimice la entropía de la partición determinada por dicho umbral [80]. Este método requiere la disponibilidad de casos de entrenamiento.

- A continuación se propone un método que combina la experiencia con los datos para llevar a cabo la discretización. El algoritmo es una modificación del método expuesto en el punto anterior. Se basa en dividir, para cada síntoma, las causas en dos grupos: las directamente relacionados con el síntoma y las no relacionadas con el síntoma. Este método, por tanto, combina el conocimiento de expertos y la información de casos de entrenamiento.

- El último método propuesto, que también combina el conocimiento con el aprendizaje a partir de datos, se basa en la teoría del test de hipótesis. El umbral de cada síntoma se calcula como el punto de cruce entre la probabilidad a posteriori de las causas relacionadas y la probabilidad a posteriori de las causas no relacionadas, dado cada uno de los síntomas. Las fdp de los síntomas dadas las causas se modelan como funciones beta, cuyos parámetros se obtienen a partir de casos de entrenamiento.

Métodos para la definición de probabilidades

En este apartado se investigan métodos para definir las tablas de probabilidad de la RB:

- De acuerdo con el primer método, son los expertos en diagnosis los que definen las tablas de probabilidad.

- El segundo método, el *Estimador de Máxima Probabilidad* (MLE), es uno de los estimadores más usados en estadística. Aproxima las probabilidades por su frecuencia relativa en los casos de entrenamiento.

- El tercer método, llamado *Estimador-m* [58], asume que las probabilidades a priori siguen distribuciones beta y, a partir de los casos de entrenamiento, calcula las probabilidades a posteriori.

- Finalmente, se propone un método que modela las fdps de los síntomas dadas las causas como funciones beta. Entonces se calculan las probabilidades de la RB como el área de las funciones betas en el intervalo correspondiente.

Métodos para combatir la imprecisión en los parámetros

A diferencia de otros campos de aplicación, como la diagnosis en medicina donde existen grandes bases de datos con casos reales [50], en diagnosis de sistemas celulares, al ser un área nueva, es difícil conseguir casos de entrenamiento. Por eso, en muchas ocasiones los modelos se basan únicamente en el conocimiento de expertos en diagnosis en redes de acceso celulares.

Debido a las dificultades para establecer los parámetros (umbrales y probabilidades) del modelo de diagnosis, se han desarrollado dos métodos que pretenden aumentar la precisión del sistema cuando los parámetros son inexactos:

- La primera técnica se ha llamado *Redes Bayesianas Suavizadas* (Smooth Bayesian Networks, SBN). El objetivo es suavizar los umbrales abruptos en la discretización de los síntomas continuos. Las SBN se pueden entender como RB a las que se ha añadido incertidumbre sobre el estado en qué se encuentran los síntomas.

- El segundo método se ha denominado *Múltiples Intervalos Uniformes* (Multiple, Uniform Intervals, MUI). Se parte de una RB en la que los síntomas han sido discretizados en sólo dos estados, para simplificar la definición de los parámetros por parte de los expertos. Entonces se añade un tercer estado alrededor del umbral, cuya probabilidad es proporcional al ancho del intervalo.

Adquisición del conocimiento

Normalmente los expertos en diagnosis no conocen las técnicas expuestas en los apartados anteriores (RB, clasificador bayesiano, etc.). Por otra parte, los expertos en inteligencia artificial no suelen tener conocimientos sobre comunicaciones móviles. Por eso, es necesario juntar ambos ámbitos de conocimiento [163, 175, 188, 76].

El objetivo de este apartado es describir un sistema que, de forma automática, construya los modelos bayesianos a partir del conocimiento de expertos en diagnosis en RANs de sistemas celulares.

En primer lugar se describe qué información debe definir el experto en diagnosis y cómo solicitar esa información en lenguaje natural de manera ordenada. Una vez que la información necesaria ha sido suministrada por el experto en diagnosis, se explica cómo construir cada uno de los modelos bayesianos descritos en este capítulo (clasificador bayesiano, SBM discreto, CIM, etc.). La información suministrada y la construcción del modelo dependen del tipo de modelo a construir.

E.4 Evaluación

La última parte de la tesis se dedica a la evaluación de los modelos y métodos para la diagnosis propuestos en esta tesis. Se expone también la metodología seguida para simular casos (con el objetivo de entrenar y probar los sistemas de diagnosis) y para analizar los modelos y técnicas desarrollados.

E.4.1 Resultados

Casos de red y simulados

Para evaluar y comparar las distintas alternativas presentadas en el Capítulo 5 se requiere un conjunto de casos de referencia que incluyan el valor de los síntomas y la causa para cada celda problemática. En este punto se resalta la dificultad de disponer de estos casos de referencia en la aplicación bajo estudio, al contrario que ocurre en otros campos como en medicina, donde existen grandes bases de datos con casos clasificados.

Si bien el sistema de gestión de las redes celulares almacena bases de datos con estadísticas sobre indicadores de funcionamiento, no se registra cuál fue la causa de la elevada tasa de llamadas caídas en una determinada celda.

Con el objetivo de obtener los casos de referencia requeridos se llevó a cabo una campaña de análisis de celdas con problemas en una red GSM/GPRS durante tres meses. En este apartado se describen los datos obtenidos, junto con sus limitaciones.

A continuación se propone un procedimiento para generar un número arbitrario de casos simulados, en base a los datos de la red real y en la aproximación de las fdp como funciones beta.

Figuras de mérito

Se describen las figuras de mérito consideradas para evaluar los distintos sistemas: precisión del diagnóstico, error en el diagnóstico, probabilidad de la causa correcta, número de orden de la causa según su probabilidad, etc.

Análisis de sensibilidad

Además de las anteriores medidas, otro parámetro importante para medir la calidad de un determinado sistema de diagnosis es la sensibilidad de sus resultados a cambios en los parámetros. Así, un sistema será tanto mejor cuanto menos sensibles sean sus conclusiones a una definición inexacta de sus parámetros. Se presenta aquí el método empírico que se utilizará para medir la sensibilidad a las probabilidades [158]. Se propone también un método para medir la sensibilidad a la imprecisión en los umbrales.

Metodología

Se describe la metodología seguida para evaluar y comparar los modelos y técnicas propuestos en la segunda parte de la tesis.

Estimación de funciones de probabilidad

En esta sección se modelan las fdp de los síntomas dadas las causas como funciones beta. Para ello se calculan los parámetros de las funciones beta utilizando un estimador de máxima verosimilitud de los casos reales obtenidos de una red GSM/GPRS. Mediante técnicas estadísticas se evalúa la proximidad entre los datos de la red y las fdp estimadas. Si los resultados son aceptables, se utilizan las funciones beta obtenidas para generar casos de prueba y casos de entrenamiento.

Evaluación de los sistemas de diagnosis

Los diversos sistemas de diagnosis presentados en el Capítulo 5 se evalúan en esta sección. Los experimentos se realizan tanto para casos de la red real como para casos simulados. Los aspectos analizados para todos los sistemas de diagnosis son las figuras de mérito, la influencia en los resultados del número de casos de entrenamiento y la sensibilidad de sus resultados a imprecisiones en los parámetros. Además para determinados métodos, como los algoritmos de discretización, se presentan ejemplos que ilustren su funcionamiento. Los sistemas evaluados han sido:

- Clasificador Bayesiano.
- Estructuras de RB: SBM y Noisy-OR.
- Aprendizaje de los parámetros del modelo, incluyendo la comparación tanto de algoritmos para la discretización de variables continuas como para el cálculo de probabilidades de la RB.
- Prevención de imprecisiones en los parámetros del modelo.

Prototipos y pruebas de campo

En esta sección se resumen algunos resultados obtenidos a raíz de la colaboración con Nokia. En primer lugar se presenta un prototipo de herramienta para la adquisición del conocimiento basada en los procedimientos descritos en el Capítulo 5. Dicho sistema se proporcionó a diversos operadores de redes de comunicaciones móviles para que construyesen un modelo de diagnosis para la red de acceso. Se explican las principales dificultades encontradas y cómo se intentaron superar. A continuación, se presenta un prototipo de herramienta automática de diagnosis. Finalmente, se describen resultados obtenidos al usar dicha herramienta de diagnosis en una campaña de pruebas en una red GSM/GPRS real. El modelo utilizado se basa en el conocimiento de expertos y fue construido con ayuda del sistema de adquisición del conocimiento.

E.4.2 Conclusiones

En este capítulo se resume el trabajo realizado a lo largo de la tesis, se describen las principales dificultades encontradas y se proponen líneas de continuación.

Las principales aportaciones de la tesis son las siguientes:

- Compilación de información, alguna conocida por expertos en diagnosis en la RAN, pero no reflejada en la literatura existente (proceso de diagnosis en redes actuales, base de conocimiento para diagnosis en GERAN, etc.).
- Utilización de técnicas previamente existentes (Clasificador Bayesiano, Redes Bayesianas, métodos de discretización, etc.) en un nuevo dominio de aplicación.
- Definición de la arquitectura de un sistema de gestión automática de fallos en redes celulares y especificación de un modelo de diagnosis para GERAN.
- Propuesta de modelos y métodos para la diagnosis automática, válidos no sólo para GSM, sino para cualquier red celular. Las principales contribuciones en este área son:
  - Modelado de las fdp de los síntomas dadas las causa como funciones beta.
  - Estructura de RB denominada Modelo de Bayes Central.

  - Método de discretización basado en minimizar la entropía, SEMD.
  - Método de discretización basado en la teoría del test de hipótesis, BMAP.
  - Método de determinación de las probabilidades de una RB basándose en funciones beta, BDF.
  - Métodos para disminuir los errores en el diagnóstico debidos a imprecisiones en los parámetros del modelo: SBN y MUI.
  - Procedimiento de adquisición del conocimiento específico para redes celulares.

- Propuesta de métodos para evaluar y comparar los sistemas de diagnosis descritos:

  - Método para generar un número arbitrario de casos, que pueden servir para evaluar sistemas de diagnosis o entrenarlos.
  - Método para evaluar empíricamente la sensibilidad a imprecisiones en la discretización de los síntomas continuos.
  - Evaluación y comparación de los sistemas de diagnosis propuestos: clasificador bayesiano, distintas estructuras de RB (SBM y Noisy-OR), algoritmos para el cálculo de probabilidades de una RB, técnicas de discretización de variables continuas y métodos para combatir las imprecisiones en los parámetros del modelo.
  - Pruebas de un sistemas de diagnosis basado en RB en una red GSM/GPRS real.

Bibliography

[1] The Celtic initiative. European institute for Research and strategic studies in telecommunications EURESCOM, www.celtic-initiative.org.

[2] www.celtic-gandalf.org.

[3] www.celtic-initiative.org/projects/gandalf/default.asp.

[4] www.eureka.be.

[5] www.moltsen.dk/products.html.

[6] www.skype.com.

[7] SYSTRA training material, 1998.

[8] Base station subsystem parameter BSSPAR. Technical Report DN98619454, Nokia Networks, Cellular Network Planning, Aug 1999.

[9] Vocabulary for 3GPP specifications. 3GPP TS GSM. Technical Specification Group Services and System Aspects. 21.905, Version 0.1.0, October 1999.

[10] Database description for BSC measurements. Technical Report DN98619454, Nokia Networks, Tampere, Finland, 2000.

[11] Fundamentals of interference in mobile networks. Application Note, June 2001.

[12] Performance management (PM); performance measurements - GSM. 3GPP TS GSM. Technical Specification Group Services and System Aspects; Telecommunication management. 32.401, Version 2.0.0, Release 4, June 2001.

[13] Physical/electrical characteristics of hierarchical digital interfaces. G.703, November 2001.

[14] Radio network optimisation course (RANOP), Nokia, August 2001.

[15] Base station system - mobile-services switching centre (BSS - MSC) interface; interface principles. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 48.002, Version 6.0.0, Release 6, December 2004.

[16] Full rate speech; comfort noise aspect for full rate speech traffic channels. 3GPP TS GSM. Technical Specification Group Services and System Aspects. 46.012, Version 6.0.0, Release 6, December 2004.

[17] Full rate speech; discontinuous transmission (DTX) for full rate speech traffic channels. 3GPP TS GSM. Technical Specification Group Services and System Aspects. 46.031, Version 6.0.0, Release 6, December 2004.

[18] Physical layer on the radio path; general description. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 45.001, Version 6.5.0, Release 6, November 2004.

[19] Signalling transport mechanism specification for the base station system - mobile-services switching centre (BSS - MSC) interface. 3GPP TS GSM, July 2004.

[20] Base station controller - base transceiver station (BSC - BTS) interface; interface principles. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 48.052, Version 6.0.0, Release 6, February 2005.

[21] Base station controller - base transceiver station (BSC - BTS) interface; layer 2 specification. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 48.056, Version 6.0.0, Release 6, February 2005.

[22] Base station controller - base transceiver station (BSC - BTS) interface; layer 3 specification. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 48.058, Version 6.1.0, Release 6, January 2005.

[23] Mobile switching centre - base station system (MSC-BSS) interface; layer 3 specification. 3GPP TS GSM, January 2005.

[24] Performance management (PM); concept and requirements. 3GPP TS GSM. Technical Specification Group Services and System Aspects; Telecommunication management. 32.401, Version 6.4.1, Release 6, February 2005.

[25] Radio network planning aspects. 3GPP TS GSM. Technical Specification Group GERAN. 45.030, Version 6.0.0, Release 6, January 2005.

[26] Radio resource control (RRC) protocol. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network; Mobile radio interface layer 3 specification. 45.018, Version 6.11.0, Release 6, January 2005.

[27] Radio subsystem link control. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 45.008, Version 6.11.0, Release 6, January 2005.

[28] Radio subsystem synchronization. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 45.010, Version 6.4.0, Release 6, January 2005.

[29] Radio transmission and reception. 3GPP TS GSM. Technical Specification Group GSM/EDGE Radio Access Network. 45.005, Version 6.8.0, Release 6, January 2005.

[30] Special issue on self-organisation in mobile networking. In C. Bettstetter, F. H. P. Fitzek, H. Hartenstein, G. Pujolle, and P. Santi, editors, *European Transactions on Telecommunications*. Wiley, oct 2005.

[31] Red.es. Ministerio de Industria, Turismo y Comercio, 2006.

[32] Wikipedia. Wikipedia Foundation, St. Petersburg, Florida, USA, http://en.wikipedia.org, 2006.

[33] Z. Altman, R. Skehill, R. Barco, L. Moltsen, R. Brennan, A. Samhat, R. Khanafer, H. Dubreil, M. Barry, and B. Solana. The Celtic Gandalf framework. In *Proc. IEEE Mediterranean Electrotechnical Conference MELECON'06*, pages 595–598, Benalmádena, Spain, May 2006.

[34] S. Andreassen, M. Woldbye, B. Falck, and S. Andersen. MUNIN: A causal probabilistic network for interpretation of electromyographic findings. In *Proc. International Joint Conference on Artificial Intelligence*, pages 366–372, Milan, Italy, August 1987.

[35] P. Antal, H. Verrelst, D. Timmerman, S. Van Huffel, B. de Moor, and I. Vergote. Bayesian networks in ovarian cancer diagnosis: Potentials and limitations. In *Proc. IEEE Symposium on Computer-Based Medical Systems*, pages 103–108, Houston, USA, June 2000.

[36] R. Barco. Knowledge Acquisition Tool specification. Technical Report AutoGERAN_KAT_2001_1H_v1_0, Nokia Networks, Málaga, Spain, June 2001.

[37] R. Barco. Conditions specification. Technical Report AutoGERAN_Cond_2002_2H_v1_0_3, Nokia Networks, Málaga, Spain, April 2002.

[38] R. Barco, F. J. Cañete, L. Díez, and V. Wille. Analysis of mobile measurement-based matrices in GSM networks. In *Proc. IEEE Vehicular Technology Conference VTC'01*, pages 1412–1416, Atlantic City, USA, October 2001.

[39] R. Barco, R. Guerrero, G. Hylander, L. Nielsen, M. Partanen, and S. Patel. Automated troubleshooting of mobile networks using bayesian networks. In *Proc. IASTED International Conference on Communication Systems and Networks (CSN'02)*, pages 105–110, Málaga, Spain, September 2002.

[40] R. Barco, P. Lázaro, L. Díez, and V. Wille. Multiple intervals versus smoothing of boundaries in the discretization of performance indicators used for diagnosis in cellular networks. *Lecture Notes in Computer Science (LNCS)*, 3483(IV):958–967, 2005.

[41] R. Barco, P. Lázaro, and J. M. Hermoso. Diagnosis de estaciones terrenas mediante redes bayesianas. In *Actas del XVII Symposium Nacional de la URSI*, pages 225–226, Alcalá de Henares, Spain, September 2002.

[42] R. Barco, P. Lázaro, V. Wille, and L. Díez. Knowledge acquisition for diagnosis in cellular networks based on bayesian networks. *Lecture Notes in Computer Science (LNCS)*, 4092, 2006.

[43] R. Barco, L. Nielsen, R. Guerrero, G. Hylander, and S. Patel. Automated troubleshooting of a mobile communication network using bayesian networks. In *Proc. IEEE International Workshop on Mobile and Wireless Communications Networks (MWCN'02)*, pages 606–610, Stockholm, Sweden, September 2002.

[44] R. Barco and R. Segura. Automated test and diagnosis tools for satellite ground stations. In *Proc. European Test and Telemetry Conference*, pages 5.34–43, Paris, France, June 1999.

[45] R. Barco, V. Wille, L. Díez, and P. Lázaro. Comparison of probabilistic models used for diagnosis in cellular networks. In *Proc. IEEE Vehicular Technology Conference VTC'06*, Melbourne, Australia, May 2006.

[46] R. Barco, V. Wille, and L. Díez. System for automated diagnosis in cellular networks based on performance indicators. *European Trans. Telecommunications*, 16(5):399–409, October 2005.

[47] G. A. Barreto, J. C. M. Mota, L. G. M. Souza, R. A. Frota, and L. Aguayo. Condition monitoring of 3G cellular networks through competitive neural models. *IEEE Trans. Neural Networks*, 16(5):1064–1075, 2005.

[48] G.A. Barreto, J.C.M. Mota, L.G.M. Souza, R.A. Frota, L. Aguayo, J.S. Yamamoto, and P. Oliveira. Competitive neural networks for fault detection and diagnosis in 3G cellular systems. *Lecture Notes in Computer Science (LNCS)*, 3124:207–213, August 2004.

[49] John Binder, Daphne Koller, Stuart J. Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2-3):213–244, 1997.

[50] C.L. Blake and C.J. Merz. UCI repository of machine learning databases. Dept. Information and Computer Science, http://www.ics.uci.edu/ mlearn/MLRepository.html, 1998.

[51] J. Breese and D. Heckerman. Topics in decision-theoretic troubleshooting: Repair and experiment. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 124–132, Portland, Oregon, August 1996.

[52] B. G. Buchanan and E. H. Shortliffe. *Rule based expert systems: The MYCIN experiments of the Stanford heuristic programming project.* Addison Wesley, MA, 1984.

[53] W. Buntine. A guide to the literature on learning graphical models. *IEEE Trans. Knowledge Data Eng.*, 8:195–210, April 1996.

[54] E.F. Castillo, J.M. Gutiérrez, and A.S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer-Verlag, New York, USA, 1997.

[55] E.F. Castillo, J.M. Gutiérrez, and A.S. Hadi. Sensitivity analysis in discrete bayesian networks. *IEEE Trans. Syst., Man, Cybern. A*, 27(4):412–423, 1997.

[56] J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Proc. European Working Session on Learning*, pages 164–178, Berlin, Germany, 1991.

[57] D. Cayrac, D. Dubois, M. Haziza, and H. Prade. Possibility theory in "fault mode effects analyses" - a satellite fault diagnosis application -. In *Proc. IEEE Intern. Conf. on Fuzzy Systems*, pages 1176–1181, Orlando, FL, June 1994.

[58] B. Cestnik. Estimating probabilities: A crucial task in machine learning. In *Proc. European Conference Artificial Intelligence (ECAI'90)*, pages 147–149, Stockholm, Sweden, August 1990.

[59] H. Chan and A. Darwiche. On the revision of probabilistic belief using uncertain evidence. In *Proc. International Joint Conference on Artificial Intelligence*, pages 99–105, Acapulco, Mexico, August 2003.

[60] D.F. Clark and A. Kandel. Fuzzy belief networks. In *Proc. ACM SIGSMALL/PC symposium on Small systems*, pages 132–142, Crystal City, USA, March 1990.

[61] J. Collins and C.E. Woodcock. Modelling the distribution of cover fraction of a geophysical field. In P. M. Atkinson and N. J. Tate, editors, *Advances in Remote Sensing and GIS Analysis*, pages 119–133. Wiley, Chichester, England, sep 1999.

[62] G.F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, pages 393–405, March 1993.

[63] G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, (4):309–348, 9 1992.

[64] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York, USA, 1999.

[65] E. Cox. *The fuzzy systems handbook*. Academic Press, 1994.

[66] P. Dagum and M. Luby. Approximating probabilistic inference in bayesian belief networks is NP-hard. *Artificial Intelligence*, pages 141–153, March 1993.

[67] B. D'Ambrosio. Local expression languages for probabilistic dependence. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 95–102, San Mateo, USA, July 1991.

[68] B. D'Ambrosio. Symbolic probabilistic inference in large BN2O networks. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 128–135, Seattle, USA, July 1994.

[69] R. Dechter and I. Rish. A scheme for approximating probabilistic inference. In *Proc. Uncertainty in Artificial Intelligence*, pages 132–141, Rhode Island, USA, August 1997.

[70] M.H. DeGroot. *Probability and statistics.* Addison-Wesley, Menlo Park, California, 1975.

[71] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, 39:398–409, 1977.

[72] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[73] R.H. Deng, A. Lazar, and W. Wang. A probabilistic approach to fault diagnosis in linear lightware networks. *IEEE J. Select. Areas Commun.*, 11:1438–1448, December 1993.

[74] P. Domingos and P.Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, November 1997.

[75] K. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Proc. International Conference on Machine Learning*, pages 194–202, Tahoe City, CA, July 1995.

[76] M. J. Druzdzel and L. C. van der Gaag. Elicitation of probabilities for belief networks: combining qualitative and quantitative information. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 141–148, Montreal, Canada, August 1995.

[77] M. J. Druzdzel and L. C. van der Gaag. Building probabilistic networks: where do the numbers come from? *IEEE Trans. Knowledge Data Eng.*, 12(4):481–486, 2000.

[78] D. Dubois and H. Prade. Fuzzy sets and probability: misunderstandings, bridges and gaps. In *Proceedings of the Second IEEE Conference on Fuzzy Systems*, pages 1059–1068, San Francisco, USA, August 1993.

[79] D. Dubois and H. Prade. Fuzzy sets - a convenient fiction for modeling vagueness and possibility. *IEEE Trans. Fuzzy Syst.*, 2(1):16–21, February 1994.

[80] U.M. Fayyad and K.B. Irani. Multi-interval discretization of continuous valued attributes for classification learning. In *Proc. International Joint Conference on Artificial Intelligence*, pages 1022–1027, Chambery, France, August 1993.

[81] J. A. Fernández, R. Barco, and P. Lázaro. Sistema de ayuda al diseño de modelos de diagnosis para redes de comunicaciones móviles. In *Actas del XX Symposium Nacional de la URSI*, Gandía, Spain, September 2005.

[82] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, November 1997.

[83] N. Friedman and M. Goldszmidt. Learning Bayesian networks with local structure. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 252–262, Portland, Oregon, August 1996.

[84] P. Frohlich, W. Nejdl, K. Jobmann, and H. Wietgrefe. Model-based alarm correlation in cellular phone networks. In *Proc. IEEE International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems*, pages 197–204, Haifa, Israel, January 1997.

[85] D. Geiger and D. Heckerman. Advances in probabilistic reasoning. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 118–126, Los Angeles, CA, July 1991.

[86] I.J. Good. *The estimation of probabilties.* M.I.T. Press, Cambridge, Massachusetts, 1965.

[87] R. Guerrero. Knowledge acquisition tool user's manual. Technical Report ATS_2H02_5, Nokia Networks, Málaga, Spain, December 2002.

[88] M. Guiagoussou and R. Boutaba. A TMN framework for faults diagnostic in wireless telecommunication networks. In *Proc. IEEE Symposium on Computers and Communications*, pages 424–433, Alexandria, Egypt, July 1997.

[89] J. F. Escobedo H. F. Assunçao and A. P. Oliveira. Modelling frequency distributions of 5 minute-averaged solar radiation indexes using beta probability functions. *Theoretical and Applied Climatology*, 75:213–224, September 2003.

[90] T. Halonen, J. Romero, and J. Melero, editors. *GSM, GPRS and EDGE Performance. Evolution Towards 3G/UMTS.* Wiley, Chichester, England, 2003.

[91] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[92] D. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196. North-Holland, Amsterdam, 1986.

[93] D. Heckerman. Causal independence for knowledge acquisition and inference. *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 122–127, July 1993.

[94] D. Heckerman. A tutorial on learning bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington, March 1995.

[95] D. Heckerman and J. Breese. A new look at causal independence. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 286–292, Seattle, Washington, July 1994.

[96] D. Heckerman, J. Breese, and K. Rommelse. Decision-theoretic troubleshooting. *Communication of the ACM*, 38(3):49–57, March 1995.

[97] D. Heckerman and J.A. Breese. Causal independence for probability assessment and inference using bayesian networks. Technical Report MSR-TR-94-08, Microsoft Research, Redmond, Washington, March 1994.

[98] D. Heckerman, J.A. Breese, and K. Rommelse. Troubleshooting under uncertainty. Technical Report MSR-TR-94-07, Microsoft Research, Redmond, Washington, January 1994.

[99] D. Heckerman, E. Horvitz, and B. Nathwani. Toward normative expert systems: Part I. the Pathfinder project. *Methods of Information in Medicine*, 31:90–105, 1992.

[100] D. Heckerman and C. Meek. Models and selection criteria for regression and classification. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 223–228, Providence, Rhode Island, August 1997.

[101] D. Heckerman, C. Meek, and G. Cooper. A bayesian approach to causal discovery. Technical Report MSR-TR-97-05, Microsoft Research, Redmond, Washington, February 1997.

[102] M. Henrion. Some practical issues in constructing belief networks. In L.N. Kanal, T.S. Leuitt, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 3, pages 161–173. Elsevier Science, Amsterdam, The Netherlands, 1989.

[103] M. Henrion, M. Pradhan, B. Del Favero, K. Huang, G. Provan, and P. O'Rorke. Why is diagnosis using belief networks insensitive to imprecision in probabilities? In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 307–314, Portland, Oregon, 1996.

[104] A. J. Hoglund, K. Hatonen, and A. S. Sorvari. A computer host-based user anomaly detection system using the self-organizing map. In *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks*, volume 5, pages 411–416, Como, Italy, July 2000.

[105] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, April 1993.

[106] A.S. Hornby and C. Ruse. *Oxford Student's Dictionary of Current English*. Oxford University Press, Oxford, UK, 1988.

[107] E. Horvitz and M. Barry. Display of information for time-critical decision making. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 296–305, Montreal, Quebec, August 1995.

[108] G. Hylander. Troubleshooting tool prototype user's guide. Technical report, Nokia Networks, Málaga, Spain, December 2001.

[109] G. Jakobson and M.D. Weissman. Alarm correlation. *IEEE Network*, pages 52–59, November 1993.

[110] F. Jensen. *Bayesian Networks and decision graphs*. Springer-Verlag, New York, USA, 2001.

[111] F. V. Jensen, U. Kjærulff, B. Kristiansen, H. Langseth, C. Skaanning, J. Vomlel, and M. Vomlelová. The SACSO methodology for troubleshooting complex systems. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 5(4):321–333, 2001.

[112] J. L. Johnson. *Probability and statistics for computer science.* Wiley, New Jersey, USA, 2003.

[113] J. L. Johnson. *Fundamental of Cellular Network Planning and Optimisation.* Wiley, West Sussex, England, 2004.

[114] R. E. Kass and L. Wasserman. Formal rules for selecting prior distributions: a review and annotated bibliography. Technical Report 583, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA, December 1993.

[115] I. Katzela and M. Schwartz. Schemes for fault identification in communication networks. *IEEE/ACM Trans. Networking*, 3(6):753–764, December 1995.

[116] R. Khanafer, L. Moltsen, H. Dubreil, Z. Altman, and R. Barco. A bayesian approach for automated troubleshooting for UMTS networks. In *Proc. 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'06)*, Karlsruhe, Germany, August 2006.

[117] J. Kim and J. Pearl. A computational model for causal and diagnostic reasoning in inference engines. In *Proc. Intern. Joint Conference on Artificial Intelligence*, pages 190–193, Karlsruhe, Germany, August 1983.

[118] O. Kipersztok and H. Wang. Another look at sensitivity of bayesian networks to imprecise probabilities. In *Proc. International Workshop on Artificial Intelligence and Statistics*, pages 307–314, Florida, USA, 2001.

[119] U. Kjærulff. Optimal decomposition of probabilistic networks by simulated annealing. *Statistics and Computing*, 2:7–17, 1992.

[120] U. Kjærulff and L. C. van der Gaag. Making sensitivity analysis computationally efficient. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 317–325, Stanford, CA, 2000.

[121] G. J. Klir. On the alleged superiority of probabilistic representation of uncertainty. *IEEE Trans. Fuzzy Syst.*, 2(1):27–31, February 1994.

[122] B. Kosko. Fuzziness vs.probability. *Int.J.General Syst.*, 17:211–240, 1990.

[123] J. Laiho, M. Kylväjä, and A. Höglund. Utilisation of advanced analysis methods in UMTS networks. In *Proc. IEEE Vehicular Technology Conference VTC'02*, pages 726–730, Birmingham, USA, May 2002.

[124] J. Laiho, K. Raivio, P. Lehtimäki, K. Hätönen, and O. Simula. Advanced analysis methods for 3G cellular networks. *IEEE Trans. Wireless Commun.*, 4(3):930–942, May 2005.

[125] J. Laiho, A. Wacker, and T. Novosad. *Radio network planning and optimisation for UMTS.* Wiley, Chichester, UK, 2002.

[126] K.B. Laskey. Sensitivity analysis for probability assessments in bayesian networks. *IEEE Trans. Syst., Man, Cybern.*, 25(6):901–909, April 1995.

[127] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201, 1995.

[128] M. Laviolette and J. W. Seaman. The efficacy of fuzzy representations of uncertainty. *IEEE Trans. Fuzzy Syst.*, 2(1):4–15, February 1994.

[129] M. Laviolette and J. W. Seaman. Unity and diversity of fuzziness-from a probability viewpoint. *IEEE Trans. Fuzzy Syst.*, 2(1):38–42, February 1994.

[130] A. M. Law and W. D. Kelton. *Simulation modeling and analysis*. McGraw-Hill, New York, 2000.

[131] P. Lázaro, R. Barco, and J.M. Hermoso. Diagnosis of earth stations using bayesian networks. In *Proc. IASTED International Conference on Artificial Intelligence and Applications*, pages 268–272, Málaga, Spain, September 2002.

[132] P. Lehtimäki and K. Raivio. A knowledge-based model for analyzing GSM network performance. In *Proc. International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, Bari, Italy, June 2005.

[133] P. Lehtimäki and K. Raivio. A SOM based approach for visualization of GSM network performance data. In *Proc. International Symposium on Intelligent Data Analysis*, Madrid, Spain, September 2005.

[134] J. Lempiäinen and M. Manninen. *Radio interface system planning for GSM/GPRS/UMTS*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.

[135] J. Lempiäinen and M. Manninen. *UMTS Radio Network Planning, Optimization and QoS Management*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[136] L. Lewis and G. Dreo. Extending trouble ticket systems to fault diagnostics. *IEEE Network*, 7(6):44–51, November 1993.

[137] D. V. Lindley. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Stat.Sci.*, 2:17–24, 1987.

[138] E. L. Madruga and L. M. R. Tarouco. Fault management tools for a cooperative and decentralized network operations environment. *IEEE J. Select. Areas Commun.*, 12(6):1121–1130, 1994.

[139] C. Melchiors and L. M. R. Tarouco. Troubleshooting network faults using past experience. In *Network Operations and Management Symposium*, pages 549–562, Honolulu, Hawaii, April 2000.

[140] T.M. Mitchell. *Machine learning*. McGraw-Hill, New York, 1997.

[141] D. C. Montgomery and G. C. Runger. *Applied statistics and probability for engineers*. Wiley, New York, 2003.

[142] M. Mouly and M.B.Pautet. *The GSM System for Mobile Communications*. Cell & Sys, Palaiseau, France, 1992.

[143] R.E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.

[144] G. Ng and K. Ong. Using a qualitative probabilistic network to explain diagnostic reasoning in an expert system for chest pain diagnosis. In *Computers in Cardiology*, pages 569–572, September 2000.

[145] T. Toftegaard Nielsen and J. Wigard. *Performance enhancement in a frequency hopping GSM network*. Kluwer academic publishers, Dordrecht, The Netherlands, 2000.

[146] D. Nikovski. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Trans. Knowledge Data Eng.*, 12(4):509–516, 2000.

[147] O. Ogunyemi, J. R. Clarke, and B. Webber. Using bayesian networks for diagnostic reasoning in penetrating injury assessment. In *Proc. IEEE Symposium on Computer-Based Medical Systems*, pages 115–120, Houston, USA, June 2000.

[148] A. O'Hagan, C.E. Buck, A. Daneshkhah, J. Richard, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements. Eliciting Experts' Probabilities*. Wiley, West Sussex, UK, 2006.

[149] K. G. Olesen, U. Kjærulff, F. Jensen, B. Falck, S. Andreassen, and S. K. Andersen. A MUNIN network for the median nerve - a case study on loops. *Applied Artificial Intelligence*, 3:384–403, 1989.

[150] G. Pagallo and D. Haussler. Boolean feature discovery in empirical learning. *Machine learning*, 5:71–99, March 1990.

[151] H. Pan. Fuzzy bayesian networks - a general formalism for representation, inference and learning with hybrid bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(7):941–963, 2000.

[152] H. Pan and D. McMichael. Fuzzy causal probabilistic networks - a new ideal and practical inference engine. In *Proc. International Conference on Multisource-Multisensor Information Fusion*, Las Vegas, USA, July 1998.

[153] H. Pan, N. Okello, D. McMichael, and M. Roughan. Fuzzy causal probabilistic networks and multisensor data fusion. In *Proc. SPIE International Symposium on Multispectral Image Processing*, volume 3543, pages 550–562, Wuhan, China, July 1998.

[154] B. Pang, D. Zhang, N. Li, and W. Kuanquan. Computerized tongue diagnosis based on bayesian networks. *IEEE Trans. Biomed. Eng.*, 51(10):1803 – 1810, 2004.

[155] A. Papoulis. *Probability, random variables, and stochastic processes.* McGraw-Hill, USA, 1984.

[156] G. Parmigiani. *Modeling in Medical Decision Making.* John Wiley & Sons, Chichester, England, 2002.

[157] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann, San Francisco, California, 1988.

[158] M. Pradhan, M. Henrion, G. Provan, B. del Favero, and K. Huang. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence*, 85(1-2):363–397, 1996.

[159] S. Qiu, A.M. Agogino, S. Song, J. Wu, and S. Sitarama. A fusion of bayesian and fuzzy analysis for print faults diagnosis. In *Proc. International Society for Computers And Their Applications*, pages 229–232, Seattle, USA, October 2001.

[160] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, March 1986.

[161] J. M. Hernando Rábanos, editor. *Comunicaciones móviles.* Centro de Estudios Ramón Areces, Madrid, Spain, 1997.

[162] J. M. Hernando Rábanos, editor. *Comunicaciones móviles GSM.* Fundación Airtel, Madrid, Spain, 1999.

[163] S. Renooij and C.L.M. Witteman. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22(3):169–194, December 1999.

[164] J.A. Rice. *Mathematical statistics and data analysis.* Wadsworth & Brooks/Cole, California, USA, 1988.

[165] I. Rish. An empirical study of the naive bayes classifier. In *Proc. International Joint Conference on Artificial Intelligence*, pages 41–46, Seattle, USA, August 2001.

[166] Stuart Russell and Peter Norvig. *Artificial intelligence : a modern approach.* Prentice Hall, New Jersey, USA, 2003.

[167] G. Salton, J. Allen, and C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108, 1994.

[168] S.L. Salzberg. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–327, September 1997.

[169] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[170] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, July 1948.

[171] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:623–656, October 1948.

[172] E. Shortliffe. *MYCIN: Computer-Based Medical Consultations.* American Elsevier, New York, USA, 1976.

[173] C. Skaanning. A knowledge acquisition tool for bayesian-network troubleshooters. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 549–557, Stanford, USA, July 2000.

[174] C. Skaanning, F. V. Jensen, and U. Kjaerulff. Printer troubleshooting using bayesian networks. *Lecture Notes in Artificial Intelligence*, 1821:367–379, 2000.

[175] C. Skaanning, F.V. Jensen, U. Kjærulff, and A.L. Madsen. Acquisition and transformation of likelihoods to conditional probabilities for bayesian networks. In *Proc. AAAI Spring Symposium on AI in Equipment Maintenance Service and Support*, pages 34–40, Palo Alto, California, March 1999.

[176] R. Skehill, M. Barry, S. McGrath, N. Gawley, M. O'Callaghan, B. González, B. Solana, R. García, L. Nielsen, M. Toril, M. Fernández, P. Lázaro, Z. Altman, H. Dubreil, R. Khanafer, R. Nasri, A. Samhat, and R. Brennan. Gandalf: Initial scenarios and project requirements. Technical Report D2.1, Celtic Telecommunication Solutions. Gandalf project, Paris, France, July 2005.

[177] C. Smith and C. Gervelis. *Cellular system design & optimization.* McGraw-Hill, 1996.

[178] S.Monti. *Learning hybrid bayesian networks from data.* PhD thesis, Univ. of Pittsburgh, Pittsburgh, July 1999.

[179] D. Spiegelhalter and S. L. Lauritzen. Sequencial updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

[180] S. Srinivas. A generalization of the noisy-OR model. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 208–215, Washington, USA, July 1993.

[181] M. Stefik. *Introduction to Knowledge Systems.* Morgan Kaufmann, San Francisco, USA, 1995.

[182] M. Steinder and A.S. Sethi. Probabilistic fault localization in communication systems using belief networks. *IEEE/ACM Trans. Networking*, 12(5):809–822, October 2004.

[183] R. Sterritt, A.H. Marshall, C.M. Shapcott, and S.I. McClean. Exploring dynamic bayesian belief networks for intelligent fault management systems. In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, volume 5, pages 3646–3652, Nashville, TN, October 2000.

[184] P. Stuckmann, Z. Altman, H. Dubreil, A. Ortega, R. Barco, M. Toril, M. Fernandez, M. Barry, S. McGrath, G. Blyth, P. Saidha, and L. Nielsen. The EUREKA Gandalf project: monitoring and self-tuning techniques for heterogeneous radio access networks. In *Proc. IEEE Vehicular Technology Conference VTC'05*, volume 4, pages 2570–2574, Stockholm, Sweden, June 2005.

[185] M. Toril, S. Pedraza, R. Ferrer, and V. Wille. Optimization of signal level thresholds in mobile networks. In *Proc. IEEE Vehicular Technology Conference VTC'02*, pages 1655–1659, Birmingham, USA, May 2002.

[186] M. Toril, V. Wille, and R. Barco. Identification of missing adjacencies in GERAN networks. *Wireless networks*, submitted.

[187] M. Valtorta, Young-Gyun Kim, and J. Vomlel. Soft evidential update for probabilistic multiagent systems. *International Journal of Approximate Reasoning*, 29(1):71–106, January 2002.

[188] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. How to elicit many probabilities. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, pages 647–654, Stockholm, Sweden, July 1999.

[189] M. Vomlelová. *Decision-theoretic troubleshooting*. PhD thesis, Univ. of Economics, Prague, Czech Republic, May 2001.

[190] C. Wang and M. Schwartz. Fault detection with multiple observers. *IEEE/ACM Trans. Networking*, 1(1):48–55, February 1993.

[191] H.R. Warner, A.F. Toronto, L.G. Veasey, and R. Stephenson. A mathematical approach to medical diagnosis - application to congenital heart disease. *Journal of the American Medical Association*, 177(3):177–183, 1961.

[192] H. Wietgrefe. Investigation and practical assessment of alarm correlation methods for the use in GSM access networks. In *Proc. IEEE/IFIP Network Operations and Management Symposium*, pages 391– 403, Florence, Italy, April 2002.

[193] H. Wietgrefe, K. Tuchs, K. Jobmann, G. Carls, P. Fröhlich, W. Nedil, and S. Steinfeld. Using neural networks for alarm correlation in cellular phone networks. In *Proc. International Workshop on Applications of Neural Networks to Telecommunications*, pages 7.2.97.1–10, Melbourne, Australia, June 1997.

[194] V. Wille, A. Kuurne, S. Burden, and R. Barco. Simulations and trial results for mobile measurement based frequency planning in GERAN networks. In *Proc. IEEE Vehicular Technology Conference VTC'02*, pages 625–628, Vancouver, Canada, September 2002.

[195] V. Wille, S. Patel, R. Barco, A. Kuurne, S. Pedraza, M. Toril, and M. Partanen. Automation and optimisation. In T. Halonen, J. Romero, and J. Melero, editors, *GSM, GPRS and EDGE performance: Evolution towards 3G/UMTS*, pages 467–511. Wiley, Chichester, England, 2003.

[196] V. Wille, M. Toril, and R. Barco. Impact of antenna downtilting on network performance in GERAN systems. *IEEE Commun. Lett.*, 9(7):598–600, July 2005.

[197] N. Wilson. Vagueness and bayesian probability. *IEEE Trans. Fuzzy Syst.*, 2(1):34–36, February 1994.

[198] R. L. Winkler. The assessment of prior distributions in bayesian analysis. *Journal of American Statistical Association*, 62:776–880, 1967.

[199] Y. Xiang, B. Pant, A. Eisen, M. P. Beddoes, and David Poole. Multiply sectioned bayesian networks for neuromuscular diagnosis. *Artificial Intelligence in Medicine*, 5(4):293–314, 1993.

[200] C.C. Yang. Fuzzy bayesian inference. In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, pages 2707–2712, Orlando, FL, October 1997.

[201] Y. Yang and G. I. Webb. A comparative study of discretization methods for naive-bayes classifiers. In *Proc. Pacific Rim Knowledge Acquisition Workshop (PKAW'02)*, pages 159–173, Tokyo, Japan, August 2002.

[202] S.A. Yemini, S. Kliger, E. Mozes, Y. Yemini, and D. Ohsie. High speed and robust event correlation. *IEEE Commun. Mag.*, 34:82–90, May 1996.

[203] S. L. Zabell. Johnson's sufficientness postulate. *The annals of statistics*, 10:1090–1099, December 1982.

[204] L. A. Zadeh. Probability theory and fuzzy logic are complementary rather than competitive. *Technometrics*, 37(3):271–276, August 1995.

[205] L.A. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8(3):338–353, 1965.

[206] L.A. Zadeh. Fuzzy logic. *Computer*, 21(4):83–93, 1988.

[207] N. L. Zhang and D. Poole. Exploiting causal independence in bayesian network inference. *Journal of artificial intelligence reasearch*, 5:301–328, December 1996.

[208] N. L. Zhang and L. Yan. Independence of causal influence and clique tree propagation. *International Journal of Approximate Reasoning*, 19:335–349, 1997.

Index